THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

Mme ESTELLE TALOUARN

le mercredi 30 septembre 2020

Titre :

Utilisation des données de séquence pour la cartographie fine et l'évaluation génomique des caractères d'intérêt des caprins laitiers français

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GENPHYSE)

Directeur(s) de Thèse :

MME CHRISTELE ROBERT-GRANIE

MME RACHEL RUPP

Rapporteurs :

M. DIDIER BOICHARD, INRA JOUY EN JOSAS

M. TOM DRUET, UNIVERSITE DE LIEGE

Membre(s) du jury :

Mme LAURENCE MOREAU, INRA MOULON, Président

M. JEROME RAOUL, INSTITUT DE L'ELEVAGE, Invité

Mme CHRISTELE ROBERT-GRANIE, INRA TOULOUSE, Membre

Mme RACHEL RUPP, INRA TOULOUSE, Membre

M. THOMAS FARAUT, INRA TOULOUSE, Membre

Remerciements

Parce que sur le plan professionnel comme personnel, on n'avance jamais seul et qu'une main tendue est toujours appréciée, je tiens à remercier toutes les personnes dont j'ai croisé le chemin au cours de mes études. Toutes ne seront pas citées dans les quelques mots qui suivent mais toutes, qu'elles soient en France ou ailleurs, ont participé à faire de moi ce que je suis aujourd'hui.

Cette thèse a été financée par la région Occitanie et le département de génétique animale d'INRAE, je tiens à les remercier pour m'avoir permis de réaliser ce travail.

Je remercie également mes deux rapporteurs Didier Boichard et Tom Druet et l'ensemble des examinateurs du jury : Thomas Faraut, Laurence Moreau et Jérôme Raoul pour avoir accepté la lourde tâche de lire ce manuscrit et d'en évaluer le contenu.

Merci à Christèle et Rachel d'avoir accepté d'encadrer cette thèse avec son lot de péripéties, ses hauts, ses bas (et ses quelques pertes de données impromptues). Merci de m'avoir fait confiance.

Merci également aux membres de mon comité de thèse : Bertrand Servin, Sophie Allais, Mekki Boussaha, Isabelle Palhière et Gwenola Tosser-Klopp pour leurs avis éclairés et leur grande force de proposition. Merci à Mekki pour son aide dans la mise en place du quality check des données de séquence. Un merci particulier à Isabelle pour sa bonne humeur et ses retours toujours constructifs sur mes analyses. Et une mention spéciale pour Gwenola qui m'a activement intégrée au Consortium caprin. Merci aussi d'avoir su arrondir les angles avec des co-auteurs du data paper pas toujours compréhensifs et parfois bien exaspérants il faut l'admettre !

Un merci tout spécial à Valérie et Flavie sans qui je ne me serais jamais engagée dans une thèse en sortie d'école d'ingénieur.

Merci à Philippe qui m’a bien aidée avec les données de séquences qui ne sont pas évidentes à prendre en main quand on n’a jamais fait de bioinformatique. Merci d’avoir pris le temps de m’aiguiller vers les bons outils !

Merci également à Florent et Julien pour avoir été aussi efficaces dans le génotypage de plus d’un petit millier d’animaux et pour avoir été compréhensifs face à mes échéances. Merci d’avoir pris le temps de tout m’expliquer.

Plus largement, merci à tous les permanents et non-permanents que j'ai pu croiser pendant tout ce temps et dont les discussions sur des sujets divers et variés sont venues enrichir ces dernières années.

Merci aux équipes GesPR et MG² pour leur accueil chaleureux.

Merci aussi au petit groupe de génomique caprine (clairement les meilleures réunions !).

Merci également aux gestionnaires d’unité qui ont géré les aspects pratiques tout en gardant le sourire !

Enfin, merci aux Stitchounettes pour avoir rendu la vie à l’INRA tellement plus gaie. :)

Merci à Laure pour le soutien dans la rédaction du data paper. Tout ça n’a pas été une mince affaire et j’ai été heureuse de partager cette expérience avec toi ! Et merci pour l’animation de ma période de confinement !

Un merci particulier à Jean-Michel, co-bureau inimitable dont le sourire, la bonne humeur et le petit accent du sud-ouest ont égayé ces 3 dernières années. Ça a été un plaisir de partager ton bureau. Merci d’avoir supporté pendant 3 ans le fait de me voir pester devant mon ordinateur ! Nos échanges variés et en particulier ceux sur les exploits tennistiques des français m’ont manqué ces derniers mois !

Un grand merci à Farid pour m’avoir fait confiance et épaulé dans l’enseignement de la génétique à l'ENSAT. Merci de m’avoir communiqué l’envie de transmettre. J’ai été plus que ravie de participer aussi activement à l’enseignement.

Je remercie également Elie, Françoise et Julien qui m'ont donné tort en me montrant qu'il était possible d'enseigner les statistiques (et merci d'avoir remis à jour mes bases dans le domaine).

Un petit mot pour tous les étudiants que j'ai pu avoir en 2 ans, merci d'avoir accepté que tout n'était pas parfait, merci pour cette réactivité en amphi et tous ces échanges. En bref, merci d'avoir rendu cette expérience valorisante et enrichissante.

Merci au Dr Rufin qui ne lira probablement pas ces quelques mots mais qui a mis fin au chemin de croix qu'a représenté mon parcours médical. Merci d'avoir fait que cette thèse s'est déroulée dans de bonnes conditions.

Une mention spéciale pour le TCCT et en particulier aux filles qui m'ont chaleureusement accueillie dans leur équipe et m'ont permis de me défouler dans la joie et la bonne humeur ! Merci pour les vaches dominicales et les apéros ! Merci notamment au comité très sélect du jeudi soir : Fanny, Séverine, Naike, Christine, Alina et Paul. Vous allez me manquer (au plaisir de se recroiser à RG ou Bercy) !

Merci aussi à Khair qui fait partie de ces rencontres inclassables dont le parcours et l'optimisme forcent le respect. Merci pour ces échanges et pour ce regard sur le monde à la fois candide et sage qui m'a tellement apporté. Merci d'avoir fait éclater ma bulle d'insouciance et de m'avoir ouvert les yeux sur des causes bien plus grandes.

Merci à Aurore, copine de galère depuis la 1ère heure du premier jour de prépa. Merci pour cet humour incomparable, ces petites escapades à droite à gauche, ces chantonnements qui me manquent tant et cette bonne humeur permanente ! Merci d'avoir toujours su me redonner le sourire en toute circonstance.

Merci à Sonia et Marie-Pierre, amies ardéchoises hors du commun qui partagent les grands moments comme les moins bons depuis tellement d'années que je ne peux même plus les compter. Les mots me manquent pour qualifier notre amitié et combien elle m'est précieuse. Trois parcours bien différents et tellement de points communs. Main dans la main.

Un merci particulier à Christian et Isabelle. Merci de m'avoir accueillie dans la famille comme si ça avait toujours été une évidence. Merci à Isabelle pour son écoute et son humanité et à Christian pour sa patience et cette capacité incroyable à transmettre sa passion du tennis.

Merci à mes grands-parents pour toutes ces valeurs transmises et pour m'avoir rappelé quotidiennement qu'un peu de recul et de mise en perspective ne fait jamais de mal à personne.

Merci à mon Papy qui, à mon grand regret, n'aura jamais vu la fin de cette thèse, lui qui n'a jamais eu accès aux études et qui les trouvait interminables, lui sans qui je ne me serais jamais intéressée à l'agriculture. Je te dois tellement. Puisque les disparus ne sont pas oubliés, tu resteras toujours présent dans un coin de ma tête comme un modèle de volonté, comme celui qui m'aura sans cesse poussée à chercher l'effort supplémentaire. J'espère t'avoir rendu fier.

Merci à ma Mamy dont le courage, l'abnégation et la capacité incroyable à tout affronter sont une grande inspiration. Merci d'être toujours présente quoiqu'il advienne.

Malgré les distances parfois colossales qui nous ont séparés au cours de mes études, merci à mes parents et mes sœurs dont l'amour et le soutien ont jalonné ma vie. Merci pour tous ces moments de bonheur simples qui forgent une vie. Merci d'avoir toujours été présents à votre façon sans jamais remettre en question mes choix et avec ce mélange frustrant de fierté et d'absence d'étonnement. Merci d'être ces exemples de force, de volonté, de courage et d'engagement qui continuent d'inspirer toutes mes décisions. En bref, merci d'incarner si bien la signification même de "Talouarn" !!

Enfin merci à celui qui a été un soutien infaillible dans toutes les épreuves et une présence quotidienne bien que plusieurs centaines de km nous aient séparés pendant plus de 3 ans. Merci d'avoir été l'instigateur d'un équilibre essentiel et le moteur de nombreux projets petits ou grands, passés ou futurs. Merci d'avoir été une source de rires, de réconfort et de bonheur tout simplement.

Table des matières

Liste des abréviations

ACP
Analyse en Composantes Principales
avpis
avant-pis
BLUP
Best Linear Unbiased Prediction
CSN1S1
Caséine alphaS1
DGAT1
Diacylglycerol O-Acyltransferase 1
DL
Déséquilibre de Liaison
DP
profondeur (depth)
DYD
Daughter Yield Deviation
EBV
Estimated Breeding Value
farrpis
forme de l'arrière-pis
ftray
forme des trayons
GBLUP
Genomic BLUP
GbS
Genotyping by Sequencing
GQ
qualité de génotype (genotype quality)
GWAS
Genome Wide Association Study
ICC
indice combiné caprin
IGGC
International Goat Genome Consortium
IMC
Index de Morphologie Caprin
indel
insertion/deletion
IPC
index de production caprin
itray
inclinaison des trayons
LSCS
Lactation average Somatic Cell Score
ltray
longueur des trayons

MAF
Minor Allele Frequency
MG
Matière Grasse
MP
Matière Protéique
NGS
Next Generation Sequencing
opied
ouverture des pieds
otray
orientation des trayons
pmam
profil de la mamelle
pplan
position du plancher
QTL
Quantitative Trait Loci
qualatarr
qualité de l'attache arrière
SNIG
Système National d'Information Génétique
SNP
Single Nucleotide Polymorphism
ssGBLUP
single-step GBLUP
TB
Taux Butyreux
TP
Taux Protéique
tpoit
tour de poitrine
WssGBLUP
Weighted sGBLUP

Liste des tableaux

Liste des figures

Introduction générale

L'espèce caprine a été domestiquée il y a environ 10 500 ans dans la région du Croissant Fertile. Depuis lors, des races se sont progressivement distinguées au sein de l'espèce, chacune s'adaptant aux conditions locales d'élevage. Les chèvres fournissent désormais du lait, de la viande et des fibres à plus grande échelle dans le monde. D'après FAOSTAT, la population caprine mondiale a augmenté de 38% depuis 1994 atteignant 1,034 milliard de têtes en 2017. La grande majorité des animaux est élevée en Asie (53,3 %) et en Afrique (40,9%), et enfin à la marge sur le continent américain (3,6%), en Europe (1,9%) et en Océanie (0,4%).

En Europe, contrairement à l'Asie ou à l'Afrique, la viande de chevreau est très peu consommée. La majorité de la production caprine est tournée vers le lait et sa transformation en fromage. La France est le 4ème pays en terme d'effectifs avec un cheptel d'environ 1,2 million d'individus en 2018. Deux pays dominent largement les effectifs européens : la Grèce avec un peu plus de 3,5 millions d'animaux et l'Espagne qui en compte 2,8 millions. Au niveau européen, on note un léger recul des effectifs depuis 2010 et la France ne fait pas exception puisqu'en 2010 le cheptel français comptait 1,4 millions d'animaux. Ce cheptel est inégalement réparti sur le territoire français : 35% des animaux sont élevés en Nouvelle Aquitaine et 14% en Occitanie, les deux grands bassins de production. Bien que les effectifs soient en baisse, la production laitière nationale est en hausse depuis 2014. La France est, par ailleurs, le premier producteur de lait en Europe et valorise la majeure partie de sa production sous forme de fromages avec de fortes valeurs ajoutées (AOP, IGP etc…). En 2017, près de 470 000 tonnes de lait ont été collectés auprès de 2 483 producteurs (source : chiffres agreste 2019).

En France, 12 races sont reconnues en tant que telles par le Ministère de l'Agriculture. Deux d'entre elles dominent très largement les effectifs du cheptel national : l'Alpine (59,1 % des effectifs au contrôle laitier en 2018) et la Saanen (37,6 %). C'est donc naturellement que l'effort de sélection se concentre sur ces deux races. Elles font l'objet d'un programme de sélection mettant en œuvre un contrôle de performance national, une gestion des accouplements ainsi qu'un centre de collecte de semence et un dispositif de testage sur descendance. Les objectifs de sélection ont sensiblement évolué au cours du temps mais restent centrés sur les caractères de production. La quantité de lait et sa composition demeurent l'objectif principal pour améliorer l'aptitude du lait à la transformation en fromage.

La conformation de la mamelle, pour favoriser l'adaptation à la traite mécanique, est entrée dans les objectifs de sélection en 1999. En 2013, c'est la qualité sanitaire du lait qui a fait son apparition dans les objectifs de sélection avec l'évaluation de la concentration de cellules somatiques dans le lait. D'autres caractères tels que la longévité fonctionnelle des chèvres sont encore à l'étude pour une mise en place prochaine (Palhière, OGET, & Rupp, 2018). En parallèle, on note un intérêt grandissant pour les caractères liés à la fertilité des femelles ou à la production de semence notamment pour la réalisation efficace de l'insémination artificielle qui constitue le principal levier de diffusion du progrès génétique.

En 2010, la première séquence du génome caprin a été assemblée par une équipe de chercheurs chinoise (Wang et al., 2013). Ainsi une femelle de race Black Yunnan a été utilisée pour fournir un premier assemblage de référence de haute qualité pour la chèvre domestique. Cet assemblage a ensuite été mis à jour (Bickhart et al., 2017). Suite à ce premier assemblage, un consortium international, l'International Goat Genome Consortium (IGGC), a été créé avec pour but de coordonner et soutenir les travaux de recherche mondiaux. Les efforts concentrés de plusieurs équipes ont abouti à la création d'une puce à ADN 50k en 2011 (Gwenola Tosser-Klopp et al., 2014). Une puce à ADN est un outil de génotypage incluant une sélection de marqueurs bi-alléliques régulièrement espacés sur le génome. Elle permet d'avoir une information synthétique sur le génome d'un individu dès sa naissance. La puce caprine a initié un ensemble de travaux utilisant des données génomiques. Ainsi des travaux de phylogénétique ou diversité des races ont été conduits (Evol et al., 2018; Mdladla, Dzomba, Huson, & Muchadeyi, 2016; Nicoloso et al., 2015; C. Oget, Servin, & Palhière, 2019; Visser, Lashmar, Marle-köster, & Poli, 2016). De même, la puce a lancé les premières analyses poussées du déterminisme génétique de caractères d'intérêt pour les éleveurs (Martin et al., 2016; Martin et al., 2018; Martin et al., 2017a; Mucha et al., 2017).

Ainsi, de nombreux QTLs (Quantitative Trait Loci) ont été localisés en races Alpine et Saanen françaises (Martin et al., 2016; Martin et al., 2017b, 2017a). Outre la région des caséines sur le chromosome 6, déjà bien connue, on peut citer : une région associée à un défaut de coloration de la robe (robe rose) en Saanen sur le chromosome 13, deux régions associées à la profondeur de la poitrine sur les chromosomes 6 et 8 et une région associée au taux butyreux sur le chromosome 8 en race Alpine. Une grande région pléiotropique a été identifiée sur le chromosome 19 en Saanen. La région du gène DGAT1 est significativement liée à la quantité de protéines et la production laitière mais aussi à la concentration des cellules somatiques dans le lait et à certains caractères de conformation de la mamelle.

En définitive, le nombre de variants causaux identifiés à partir des informations génotypiques de la puce 50k reste faible. Seules deux mutations dans la région DGAT1 (Diacylglycerol O-Acyltransferase 1) ont été pleinement caractérisées. La région des caséines, bien que petite (environ 0,233 Mb), contient de nombreux polymorphismes qui rendent sa caractérisation compliquée. Pour la caséine alphaS1 (gène CSN1S1), c'est la combinaison des génotypes de plusieurs substitutions et insertions/délétions qui permet d'établir le génotype de l'individu. En l'état actuel toutefois, les génotypages 50k ne permettent pas de prédire avec certitude le génotype d'un individu pour CSN1S1 et l'analyse de référence faisant notamment appel à des enzymes de restriction reste nécessaire.

La puce a de plus permis d'instaurer les premières évaluations génomiques caprines, officiellement mises en place en 2018 pour les deux principales races (Carillier-Jacquin et al., 2016; Carillier et al., 2013; Teissier et al., 2018). La sélection génomique s'appuie sur le génotypage et phénotypage d'une grande population constituée, à ce jour, pour la majorité de mâles. Pour obtenir des performances, ces mâles sont testés sur descendance. Les animaux génotypés et phénotypés servent de référence pour établir un lien statistique entre génotypes aux marqueurs et phénotype. Une fois l'équation de prédiction des phénotypes établie, elle sert à prédire le potentiel génétique des animaux candidats dès la naissance, uniquement à partir de leur génotype. Une première sélection des individus peut ainsi être effectuée très tôt et réduit par conséquent le coût d'élevage des candidats à la sélection et l'intervalle de génération. Il est toutefois nécessaire d'actualiser régulièrement l'équation de prédiction en ajoutant des individus à la population de référence.

Dans le cas de génotypes obtenus par une puce à ADN, la fiabilité de la prédiction des phénotypes dépend de la force du lien entre les marqueurs de la puce et la mutation causale responsable des variations phénotypiques observées. On parle de déséquilibre de liaison (DL). Le déséquilibre de liaison se définit généralement comme l'association non-aléatoire de deux allèles de deux marqueurs ou gènes différents. Les marqueurs associés par déséquilibre de liaison sont en général proches sur le génome et leurs allèles présentent des fréquences similaires dans la population. Plus la distance entre le marqueur et la mutation est grande plus la possibilité qu'une recombinaison entre les deux ait lieu est importante (et donc plus le DL entre le marqueur et la mutation est faible). Ainsi, l'intégration directe de données de séquence, qui contiennent par définition les mutations causales, dans les équations de prédiction permettrait de s'affranchir complètement du degré de liaison entre le marqueur et la mutation causale. Elles devraient donc accroître la précision des évaluations génomiques.

Les données de puces à ADN bien qu'intéressantes du point de vue du coût ont montré leurs limites quant à l'identification de candidats fonctionnels et la mise en place d'évaluations génomiques précises. Désormais, de nombreuses espèces se tournent vers le séquençage NGS (Next Generation Sequencing) pour affiner la recherche de QTL et intégrer ces résultats dans les évaluations génomiques. En effet, la séquence permet d'accéder à l'intégralité de la variabilité génétique d'un individu et donc potentiellement aux mutations causales d'un phénotype particulier. La récente diminution des coûts de séquençage permet d'envisager un séquençage plus massif de l'espèce caprine. Le jeu de données de séquence initial ayant servi à la création de la puce est limité (5 gene pools : Alpine, Boer, Creole, Katjang/Savanna and Saanen) et ne représente pas la diversité du genre *Capra* que l'on peut observer dans le monde. Le projet VarGoats soutenu par l'IGGC a pour but de fournir un jeu de données de séquence plus conséquent (N > 1000) qui représente au mieux la variabilité génétique totale de l'espèce *Capra hircus* tout en intégrant des individus sauvages du genre *Capra*. Les races déjà séquencées en première intention dans le jeu initial ont également bénéficié de séquences supplémentaires. Ainsi, parmi les individus séquencés dans le cadre du projet, on compte désormais 44 Alpines et 37 Saanen. Ce nombre est insuffisant pour conduire directement des analyses d'association, il permet néanmoins d'envisager une imputation de l'information depuis des typages 50k vers la séquence.

Les travaux de cette thèse exploitent les données haut-débit (séquences) récemment acquises en Alpine et Saanen françaises. De nouvelles régions associées à des caractères d'intérêt pourront être mises à jour et les zones précédemment identifiées sur les typages 50k pourront être affinées. L'étude de différents scénarii d'imputation et de leur faisabilité constitue un préalable nécessaire à toute analyse d'association sur séquence. Une fois les typages 50k disponibles imputés, la détection de régions d'intérêt devrait permettre d'identifier des candidats qui pourront potentiellement être inclus dans les évaluations génomiques. L'impact de l'intégration de données de séquence sur la précision des évaluations génomiques sera lui-même quantifié.

Les travaux menés dans le cadre de cette thèse seront exposés après une synthèse bibliographique qui éclairera le lecteur sur les enjeux des recherches qui ont été conduites pendant trois ans. Dans une deuxième partie, nous détaillerons les données de séquence disponibles, leur traitement ainsi que le processus d'imputation vers la séquence en caprins laitiers français. Puis, dans un troisième temps, nous nous attacherons plus particulièrement à la cartographie fine de la région QTL pléiotrope du chromosome 19 dans la race Saanen en utilisant notamment les données de séquences disponibles. Dans le quatrième chapitre, nous

aborderons comment les résultats obtenus précédemment ont pu bénéficier aux évaluations génomiques. Enfin dans un dernier temps, nous discuterons de l'ensemble des travaux et des perspectives à donner à ces travaux.

Chapitre 1
Synthèse bibliographique

Dans ce chapitre, nous détaillerons en premier lieu l'organisation et les applications de la sélection dans la filière caprine en France. Nous évoquerons ensuite l'évolution récente des méthodes de séquençage, la gestion de ces nouvelles données et comment elles peuvent être intégrées dans différentes analyses : l'imputation, les analyses d'association et les évaluations génomiques.

I. La sélection dans la filière caprine laitière française

I.1. Importance de la production laitière caprine

Bien que l'élevage caprin ne représente qu'une très faible part du produit intérieur brut national, la France est un producteur de lait de chèvre majeur en Europe. Au recensement agricole de 2010, on comptait 7 600 exploitations professionnelles avec plus de 10 femelles. Parmi celles-ci, 3 000 exploitants sont considérés comme livreurs stricts, c'est-à-dire qu'ils ne produisent que du lait qui sera par la suite transformé en fromage à l'extérieur de l'exploitation, 2 900 exploitants transforment le lait produit en fromage et enfin 1 700 exploitants ne produisent pas de lait (viande de chevreau, écopâturage etc…). D'après une enquête du Service de la Statistique et de la Prospective en 2013, les livreurs détiennent 72% du troupeau national avec un cheptel moyen d'environ 237 individus. D'après l'agreste, en 2016, ce sont près de 603 millions de litres de lait qui sont produits en France dont 77% sont livrés à l'industrie. Cette production est relativement concentrée dans les régions d'une grande moitié Ouest du territoire (Figure 1).

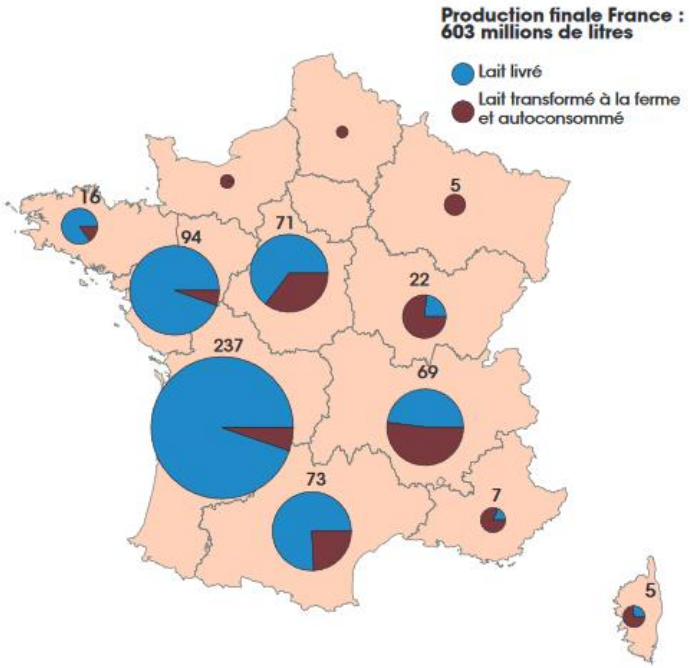


Figure 1: Répartition de la production laitière caprine en France
(source : Institut de l'Elevage)

Douze races sont reconnues par le Ministère de l'Agriculture : la Corse, la Créole, la chèvre des Fossés, la chèvre du Massif Central, l'Angora, l'Alpine, la Lorraine, la Poitevine, la Provençale, la Pyrénéenne, la Rove et la Saanen. Les races laitières Alpine et Saanen représentent l'écrasante majorité des effectifs du troupeau français. Ces deux races trouvent une origine commune dans une race du Nord des Alpes. En 2018, Alpine et Saanen représentaient 59,1% et 37,6% des lactations enregistrées sur la campagne 2018 du contrôle laitier. Une sélection efficace a permis à la France d'être le premier pays producteur de lait avec seulement le 4ème troupeau d'Europe. Lors de la campagne de 2018, l'Alpine (Figure 2) était en mesure de produire en moyenne 942 kg de lait sur une durée de 313 jours de lactation. Les taux butyreux (TB) et protéique (TP) atteignaient respectivement 37,8 et 33,5 g/kg dans la race. La Saanen (Figure 3) présentait une lactation légèrement plus longue avec 330 jours pour produire en moyenne 1 010 kg de lait avec des taux un peu inférieurs à l'Alpine : 36,2 pour le TB et 32,4 pour le TP.

Figure 2: Chèvre de race Alpine
(source: Soignon)

Figure 3: Chèvre de race Saanen
(source: Wikipedia)

I.2. Organisation de la sélection et acteurs

Aujourd'hui, Capgènes (www.capgenes.com/) est l'organisme en charge de la sélection des caprins en France. L'organisation dans sa forme actuelle est née en 2008 et regroupe 13 coopératives ou unions de coopératives. Elle est présentement responsable de la gestion de trois schémas de sélection génétique : un pour chacune des grandes races laitières (Alpine et Saanen) et enfin un en race Angora qui est élevée pour la production de mohair. L'entreprise est aussi en charge de la conservation des autres races caprines reconnues par le Ministère de l'Agriculture.

Capgènes est responsable de la diffusion du progrès génétique, l'entreprise est également l'unique centre français de collecte de semence en caprin et produit près des 300 000 doses par an dont 20 000 sont exportées à l'international. Les paillettes servant aux inséminations artificielles y sont produites et utilisées pour diffuser le progrès génétique à l'ensemble de la filière. Le recours à l'insémination reste minoritaire dans la filière (environ 9% des chèvres) ce qui implique entre autres des problèmes d'affiliation paternelle des femelles. D'après Capgènes qui a recensé, dans le cadre du projet Gènes Avenir, les déclarations de saillies permettant l'affiliation paternelle des femelles, près de 45% des femelles étaient sans affiliation paternelle en octobre 2018. Toutefois, le nombre de déclarations de saillies était en augmentation en 2017 atteignant près de 172 000.

L'insémination artificielle est aussi utilisée dans le cadre de plans d'accouplements programmés, en particulier pour l'insémination des mères à boucs (4% des meilleures femelles de la base de sélection). Le choix de ces dernières est effectué dans une optique d'amélioration génétique et de gestion de la diversité intra-race (limitation de la consanguinité à 2%) (Colleau, Moureaux, Briend, & Bechu, 2004). Chaque année, 1 000 accouplements sont programmés. Environ 200 mâles issus de ces accouplements entrent en quarantaine dans le centre de sélection. Ils sont en premier lieu retenus ou non sur la base de leur croissance, leur conformité au standard de la race (notes de pointages), leur état sanitaire et après un premier examen de leur appareil génital. Ces premières vérifications conduisent à la sélection de 120 boucs. Enfin 70 mâles sont sélectionnés pour entrer en centre de production de semence en fonction de leur comportement sexuel (capacité à sauter etc…) et leur aptitude à produire de la semence de qualité (volume et concentration de l'éjaculat, motilité et anomalie des spermatozoïdes, aptitude à la congélation/décongélation). Ces mâles seront évalués sur la base d'un testage sur descendance. La Figure 4 présente l'organisation de la sélection avec les effectifs qui seront mis en place dans les prochaines années. Le testage sur descendance est un dispositif conséquent par lequel environ 200 inséminations artificielles (IA) sont effectuées pour chaque candidat. Entre 80 et 100 filles pour chacun des mâles en testage sont ensuite soumises au contrôle de performances par an. Les boucs finalement agréés sont enfin sélectionnés sur la base de leur valeur génétique. Ils servent d'améliorateurs de la race et peuvent avoir un nombre important de filles réparties sur tout le territoire français (jusqu'à 2 000 filles). Les boucs de renouvellement qui remplaceront ces boucs en fin de carrière, sont issus de l'accouplement programmé de boucs améliorateurs aux meilleures femelles ayant au minimum 4 lactations enregistrées au contrôle de performance.

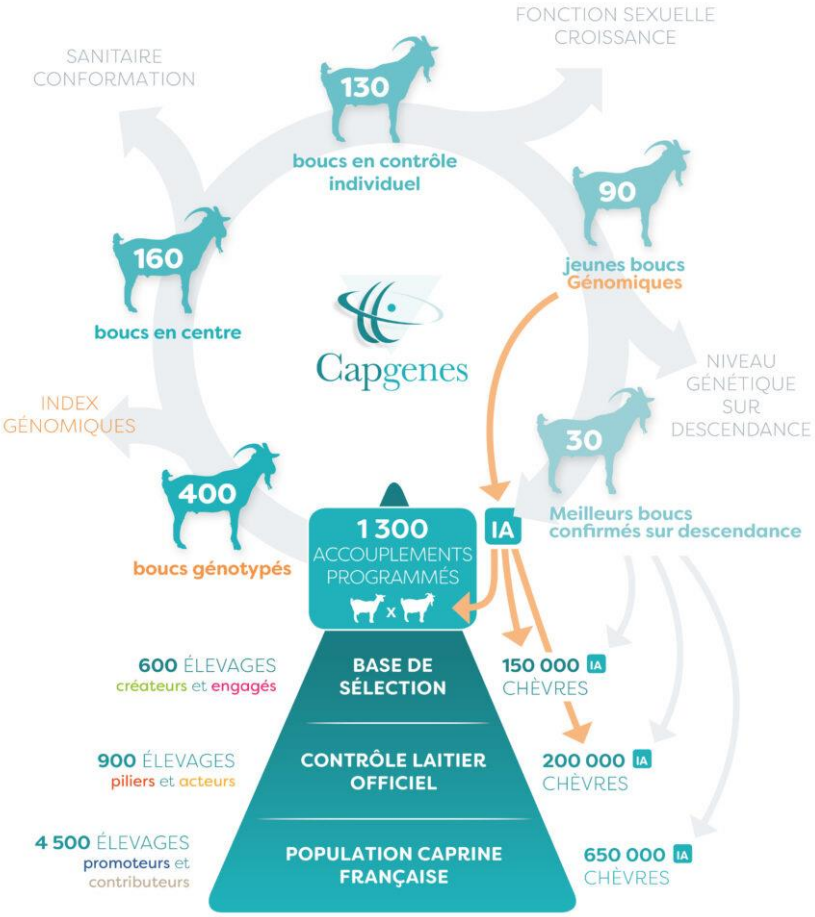


Figure 4: Schéma de sélection des races laitières françaises Alpine et Saanen
(source :https://www.capgenes.com/)

I.3. Le contrôle de performance et les caractères d'intérêt

Parmi les éleveurs laitiers, 1 487 étaient adhérents au contrôle laitier lors de la campagne 2018 ce qui représente environ 19,5 % des éleveurs. L'adhésion au contrôle permet à l'éleveur d'obtenir un bilan sur son troupeau chaque année en fin de campagne, il peut ainsi se situer par rapport au reste du cheptel français. Les performances des femelles sont relevées mensuellement par des techniciens et alimentent une base de données nationale du SNIG (Système National d'Information Génétique) qui sert ensuite à l'indexation. Pour chaque femelle, des paramètres de production sont ainsi mesurés régulièrement : la quantité de lait, les quantités de matières protéique (MP) et de matière grasse (MG), les taux butyreux (TB) et protéique (TP).

La santé de la mamelle est, quant à elle, mesurée grâce aux comptages du nombre de cellules somatiques dans le lait (CCS) dans le cadre du contrôle laitier. La présence de ces cellules immunitaires est un indicateur du niveau d'inflammation de la mamelle. Ces inflammations, ou mammites, sont majoritairement d'origine infectieuse (Rupp et al., 2019). Par définition, la variable CCS n'est pas normalement distribuée, elle subit donc une transformation logarithmique selon la formule suivante : $SCS = Log_2\left(\frac{CCS}{100\,000}\right) + 3$. Les différents SCS obtenus au cours d'une même lactation sont ensuite corrigés pour les effets du rang et du stade de lactation. Ils sont enfin pondérés et moyennés pour obtenir une valeur unique pour la lactation, on parle alors de LSCS (Lactation average Somatic Cell Score).

Enfin, les caractères de morphologie de la mamelle (Manfredi, Piacere, Lahaye, & Ducrocq, 2001) sont pointés par les techniciens de Capgènes lors de la première lactation de chaque femelle au contrôle (Figure 5). Parmi ces caractères, on compte le tour de poitrine, la forme de l'avant-pis, la forme de l'arrière-pis, la forme, longueur, orientation et inclinaison des trayons, la qualité de l'attache arrière, la position du plancher ou distance plancher-jarret, le profil de la mamelle et l'ouverture des pieds. Les caractères de conformation sont notés sur une échelle allant de 1 et 9 (tour de poitrine et longueur des trayons exceptés). En 2017, le pointage a été effectué par 11 pointeurs agréés qui ont visité près de 580 élevages dans 61 départements et pointé environ 37 000 femelles.

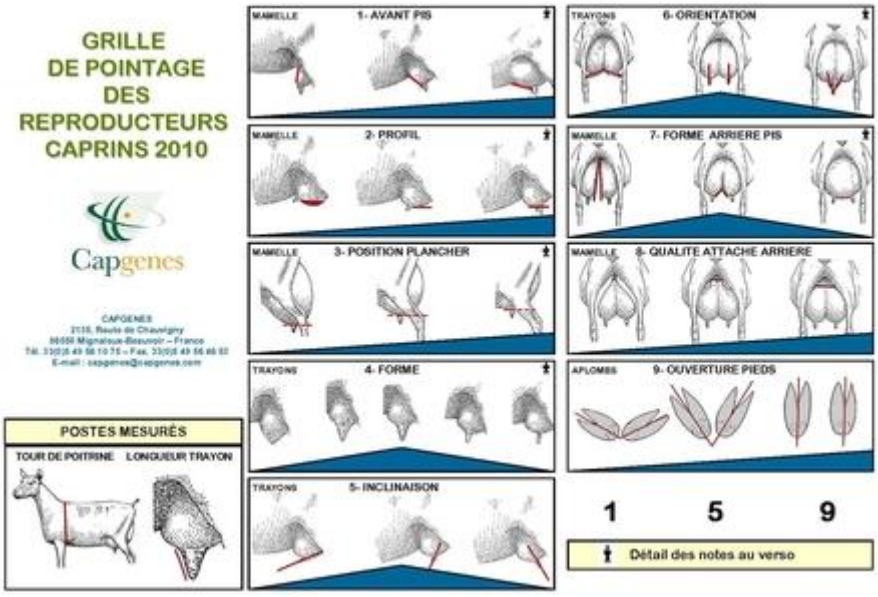


Figure 5: Grille de pointage des reproducteurs pour les caractères de morphologie
(source : CapGenes)

I.4. L'objectif de sélection en caprins laitiers français

L'efficacité de la sélection sur un caractère est mesurée par le progrès génétique (ΔG). Le progrès génétique se définit très généralement comme l'augmentation du niveau génétique par an. Il suit la formule suivante : $\Delta G = \frac{\mathrm{iR}\sigma_G}{\mathrm{T}}$. Cette formule reste toutefois très générale et suppose que tous les individus dans la population sont sélectionnés de la même façon. Ce n'est pas le cas dans les filières animales, il existe plusieurs voies de sélection père-fils, père-fille, mère-fille et mère-fils.

R correspond à la précision de l'évaluation génétique. Elle se définit comme la corrélation entre la valeur génétique vraie et la valeur génétique estimée. En réalité, la valeur génétique vraie n'est pas accessible, R est donc approchée par la racine du coefficient de détermination (CD) qui est un indice de fiabilité de l'évaluation compris entre 0 et 1. L'intensité de sélection (i), exprimée en écart-type du critère utilisé, représente l'effort de sélection, elle correspond à l'écart entre la valeur génétique moyenne des individus sélectionnés et celles des candidats à la sélection.

L'intervalle de génération (T) est l'âge moyen des parents à la naissance de leurs descendants. Une augmentation de cet âge réduit le progrès génétique annuel. Ce paramètre dépend en premier lieu de la biologie de l'espèce en question et en particulier de l'âge auquel les animaux sont en capacité d'avoir des descendants. Néanmoins, une organisation différente

de la sélection peut réduire cet intervalle. En effet, avant l'arrivée de puces de génotypage, les candidats étaient uniquement évalués avec les performances de leurs filles. Il était donc nécessaire d'attendre qu'un candidat ait des filles et que ces dernières soient entrées en production. Bien que le testage sur descendance soit maintenu en caprins, les puces de génotypage permettent d'obtenir une valeur génétique estimée dès la naissance d'un individu. Le récent passage à la génomique a donc permis de réduire le temps nécessaire à l'obtention d'une valeur génétique tout en conservant une bonne fiabilité des évaluations augmentant de fait le progrès génétique réalisable par an.

Enfin, ΔG dépend fortement de l'écart-type génétique du caractère (σ_G). Ce dernier mesure la variabilité génétique du caractère dans la population. Plus cette diversité est grande plus le progrès génétique possible est important. Un corollaire à l'écart-type génétique est l'héritabilité du caractère (h^2). L'héritabilité est la part de la variabilité phénotypique (σ_P^2), c'est-à-dire observable dans la population étudiée, qui est explicable par la variabilité génétique additive de la population : $h^2 = \frac{\sigma_G^2}{\sigma_P^2}$. Pour qu'un caractère soit sélectionnable, il faut qu'il ait une héritabilité raisonnable et/ou une variabilité génétique suffisante. En caprin, l'héritabilité des caractères en sélection est variable (Tableau 1) (Clément et al., 2006; Manfredi & Ådnøy, 2012; Rupp et al., 2011). Il est à noter que les héritabilités des différents caractères sont similaires dans les deux races Alpine et Saanen. Les héritabilités les plus élevées sont observées pour les taux. Le caractère de santé de la mamelle (LSCS) est celui qui est le moins héritable parmi les caractères pris en compte en sélection. On peut donc espérer un progrès moindre sur la résistance aux mammites que sur les autres caractères.

Tableau 1: Héritabilité de différents caractères en caprins laitiers français (Alpine et Saanen) (Clément et al., 2006; Teissier, 2019)

	ALPINE	***SAANEN***
LAIT	0,32	0,34
MG	0,37	0,40
MP	0,36	0,34
TB	0,58	0,60
TP	0,58	0,50
LSCS	0,19	0,21
AVANT-PIS	0,32	0,29
FORME DE L'ARRIERE-PIS	0,31	0,23
FORME DES TRAYONS	0,31	0,30
INCLINAISON DES TRAYONS	0,18	0,18
LONGUEUR DES TRAYONS	0,45	0,42
OUVERTURE DES PIEDS	0,14	0,13
ORIENTATION DES TRAYONS	0,33	0,25
PROFIL DE LA MAMELLE	0,37	0,25
POSITION DU PLANCHER	0,31	0,35
QUALITE DE L'ATTACHE ARRIERE	0,27	0,29
TOUR DE POITRINE	0,51	0,48

L'amélioration génétique des caractères en sélection peut être rendue complexe par l'interaction de ces caractères entre eux. Les corrélations génétiques entre les caractères sont donc systématiquement calculées avant l'intégration d'un nouveau caractère en sélection. Il est souvent nécessaire de trouver un compromis pour intégrer un caractère sans en dégrader un autre. Ainsi, sélectionner pour la production de lait améliore simultanément les matières grasse et protéique produites pendant la lactation mais dégrade potentiellement les taux par un phénomène de dilution (Tableau 2) (Manfredi & Ådnøy, 2012). De même, pour les caractères de morphologie (Tableau 3) (Clément et al., 2006; Manfredi et al., 2001), sélectionner pour le caractère de profil de la mamelle (pmam) améliore indirectement le caractère d'orientation des trayons (otray) alors que ce caractère (pmam) a beaucoup moins d'influence sur la sélection pour la forme de l'arrière-pis (farrpis) ou la position du plancher (pplan). La corrélation entre les caractères de production et le gabarit de l'animal (mesuré à travers le tour de poitrine) est faible (inférieure à 0.1) (Manfredi et al., 2001). Excepté pour les caractères de conformation de la mamelle qui relèvent de la suspension de cette dernière, la corrélation avec la production laitière s'est avérée inférieure à 0,2 et 0,1 en Alpine et Saanen respectivement

(Manfredi et al., 2001). Enfin les corrélations entre le LSCS et les caractères de production laitière sont inférieures à 0,20 dans les 2 races. Les corrélations entre LSCS et caractères de morphologie sont elles aussi faibles (Rupp et al., 2011) : comprises entre -0,24 et 0,34 en Alpine et entre -0,19 et 0,15 en Saanen (Rupp et al., 2011).

Tableau 2: Corrélations génétiques des caractères de production laitières en sélection en Alpine et Saanen (Clément et al., 2006; Teissier, 2019)

	LAIT	***MG***	***MP***	***TB***	***TP***
LAIT	1	0,85/0,86	0,93/0,95	-0,16/-0,12	-0,38/-0,40
MG		1	0,88/0,88	0,38/0,39	-0,11/-0,16
MP			1	0,01/-0,01	-0,04/-0,11
TB				1	0,49/0,41

abréviations : MG : Matière Grasse ; MP : Matière Protéique ; TB : Taux Butyreux ; TP : Taux Protéique

Tableau 3: Corrélations génétiques des caractères de morphologie de la mamelle pointés en Alpine et Saanen (Clément et al., 2006; Teissier, 2019)

	AVPIS*	***FARRPIS***	***FTRAY***	***ITRAY***	***LTRAY***	***OTRAY****	***PMAM****	***PPLAN****	***QUALATARR****
AVPIS	1	0,23/0,32	0,39/0,38	0,50/0,23	-0,28/-0,41	0,11/0,17	-0,03/-0,17	0,55/0,53	0,48/0,49
FARRPIS		1					-0,54/-0,39	0,10/0,22	-0,02/0,12
FTRAY			1	0,01/-0,02	-0,50/-0,47	-0,19/-0,17	-0,58/-0,61	0,29/0,30	0,27/0,15
ITRAY				1	0,08/-0,06	0,26/0,22	0,36/0,37	0,08/0,16	0,26/0,21
LTRAY					1	0,40/0,30	0,70/0,67	-0,35/-0,30	-0,26/-0,18
OTRAY						1	0,63/0,61	0,17/0,12	0,37/0,25
PMAM							1	0,14/0,09	0,19/0,19
PPLAN								1	0,71/0,74
QUALATARR									1

*utilisé dans le calcul de l'index morphologie (IMC)

abréviations : avpis : avant-pis ; farrpis : forme de l'arrière pis ; ftray : forme des trayons ; itray : inclinaison des trayons ; ltray : longueur des trayons ; opied : ouverture des pieds ; otray : orientation des trayons ; pmam : profil de la mamelle ; pplan : position du plancher ; qualatarr : qualité de l'attache arrière ; tpoit : tour de poitrine

Comme dans les autres filières de ruminants, le programme de sélection se focalise sur la voie mâle qui permet de diffuser plus facilement le matériel génétique sur un grand nombre d'élevages. Initialement seuls le taux protéique et la quantité de protéines du lait étaient pris en compte dans l'objectif de sélection. Depuis 1999, il inclut la matière grasse du lait (quantité et taux). Cette démarche vise à améliorer la fromageabilité du lait et a donc un poids important dans l'objectif de par son importance économique. Des caractères liés à la conformation de la mamelle (profil de la mamelle, hauteur du plancher, qualité de l'attache arrière, orientation des trayons et forme de l'avant-pis) ont été ajoutés dans un caractère synthétique en 2006. La résistance aux mammites par l'intermédiaire du comptage des cellules somatiques est intégrée dans le choix des boucs d'insémination mais n'a toutefois pas encore été intégré dans l'index synthétique des animaux. Un index spécifique sur ce caractère est cependant fourni aux éleveurs (Clément et al., 2015). Des caractères fonctionnels tels que la longévité fonctionnelle des femelles, la fertilité des femelles et la production de semence chez les mâles ne sont pas encore intégrés dans l'objectif bien qu'ils soient ou aient été étudiés.

La sélection d'individus reproducteurs repose sur le calcul d'index de synthèse qui combine plusieurs caractères dont l'importance est pondérée par leur valeur économique. Ainsi, l'indice combiné caprin (ICC) résume deux autres index : l'index de production caprin (IPC) et l'index de morphologie caprin (IMC). En Saanen, l'ICC est calculé comme ceci : ICC= IPC + 0,6IMC alors qu'en Alpine une importance légèrement moindre est donnée à la morphologie : ICC = IPC + 0,5IMC. La formule de l'IPC est identique dans les deux races, il est calculé à partir des index des caractères laitiers : IPC = 0,59 Index MP + 0,26 Index TP + 0,12 Index MG + 0,06 Index TB. L'IMC regroupe, quant à lui, les caractères de morphologie de la mamelle sous la formule : IMC = 0,2 Index pmam + 0,2 Index pplan + 0,2 Index qualatarr + 0,2 Index avpis + 0,2 Index otray. Pour des raisons pratiques, la formule de l'ICC n'a pas été modifiée pour intégrer le comptage des cellules somatiques. Pour faciliter la diffusion aux éleveurs ce dernier est fourni à part. Les parts relatives des différents index ont été réestimées en Alpine et Saanen et sont présentées sur la Figure 6.

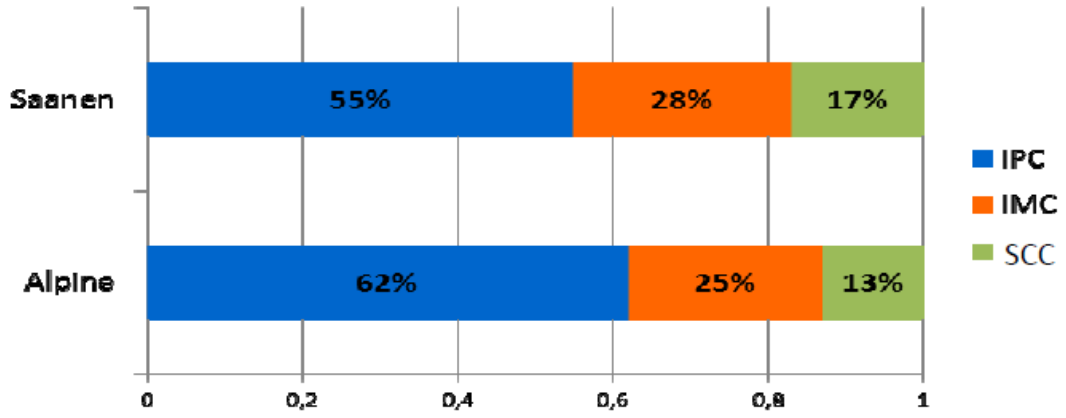


Figure 6: Part de chacun des index synthétiques et des cellules dans l'objectif de sélection en caprins laitiers français en 2020
(source: Virginie Clément, Institut de l'Elevage, communications personnelles)

En 2014, le schéma de sélection des reproducteurs basé sur la méthode de parenté minimum a été évalué sur une période s'étendant de 2006 à 2013. Cette évaluation a conduit notamment à une estimation du progrès génétique atteint. Ainsi Palhiere et al. (2014) ont évalué que l'IPC qui combine les différents caractères de production a progressé sur la période d'environ 0,15 et 0,20 écart-type génétique par an en Alpine et Saanen respectivement. Cette progression s'est faite malgré l'implémentation d'une sélection sur la morphologie de la mamelle (pourtant négativement corrélée). L'IMC a, quant à lui, très peu progressé sur la même période ce qui peut être dû à sa récente mise en place et à son plus faible poids dans l'index de synthèse ICC. En 2017, les progrès annuels ont été réévalués, les chiffres obtenus sur la période 1996-2018 sont présentés sur la Figure 7.

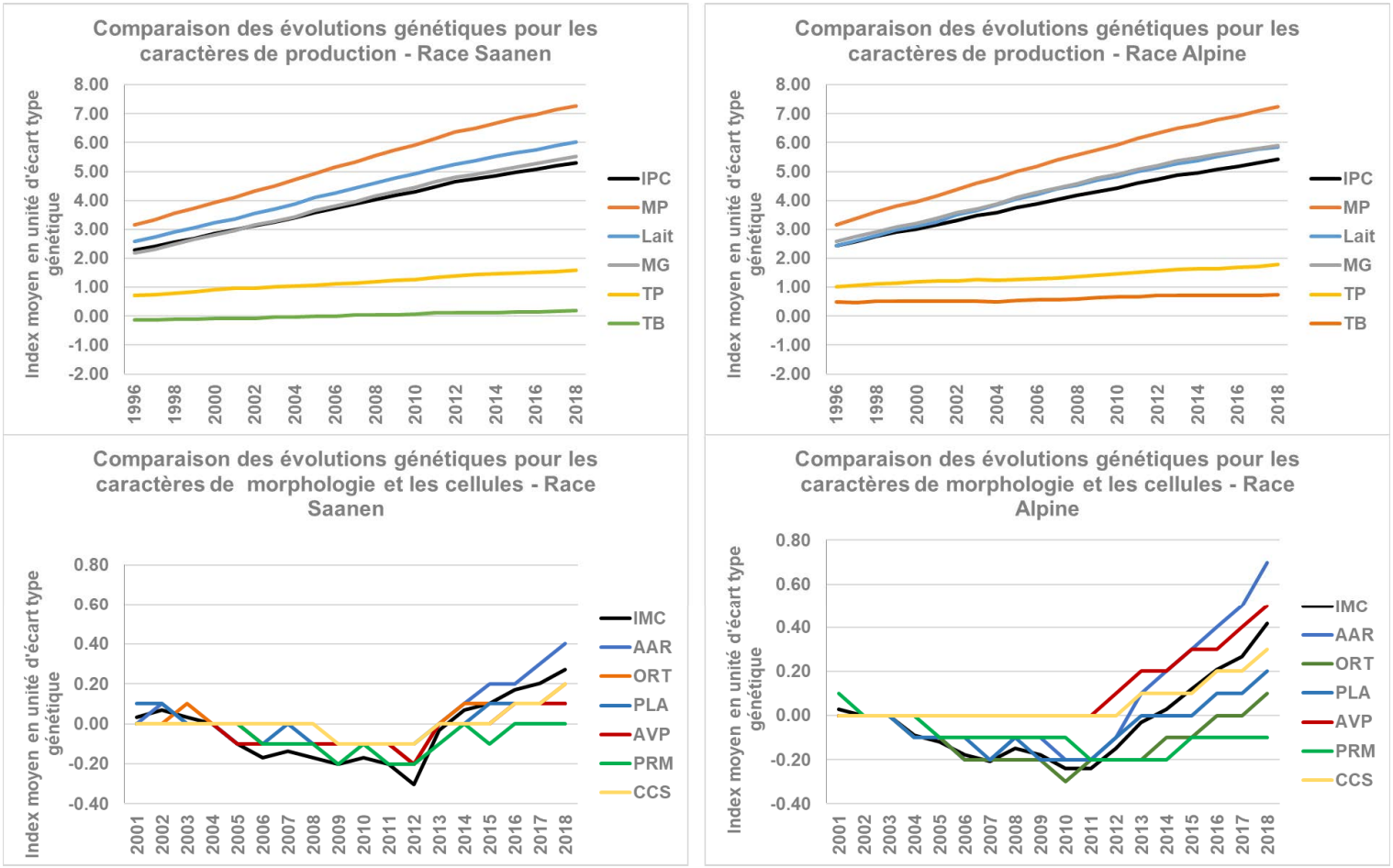


Figure 7: Evolution des différents caractères pris en compte en sélection chez les caprins laitiers français sur la période 1996-2018

abréviations : IPC : Index de Production Caprin ; MP : Matière Protéique ; MG : Matière Grasse ; TP : Taux Protéique ; TB : Taux Butyreux ; IMC : Index de Morphologie Caprin ; AAR : Qualité de l'attache-arrière ; ORT : Orientation des trayons ; PLA : Position du Plancher ; AVP : Avant-pis ; PRM : Profil de la mamelle ; CCS : Comptages de Cellules Somatiques

II. L'ADN : support de l'information génétique

II.1. La structure moléculaire de l'ADN

L'ADN (Acide DésoxyriboNucléique) est le support universel de l'information génétique dans le monde du Vivant. Cette molécule est constituée de deux chaînes orientées, complémentaires et reliées entre elles par des liaisons hydrogènes. Ces deux chaînes sont des polymères de nucléotides (Figure 8). Un nucléotide est l'assemblage d'un phosphate avec un sucre (désoxyribose) et une base azotée. Cette base peut être de 2 types : une purine (adénine ou guanine) ou une pyrimidine (thymine ou cytosine). La séquence de bases azotées définit le code génétique d'un individu. Cette séquence comporte notamment des gènes (environ 2% de la séquence) qui sont des unités fonctionnelles. Un gène comporte des régions non-traduites, les introns, des régions transcrites, les exons, et des régions régulatrices de son expression (Figure 9). Ces derniers donneront dans certains cas une protéine. Les gènes sont lus et interprétés d'un bout à l'autre par une fenêtre non-chevauchante de 3 nucléotides ou codons. C'est la succession des codons qui donne la séquence des acides aminés de la protéine correspondante suivant un tableau de correspondance précis (Figure 10).

L'intervention d'une famille de protéines appelées histones permet de compacter progressivement la double hélice d'ADN par interaction successive avec l'ADN puis entre les histones elles-mêmes. La forme la plus compactée est le chromosome (Figure 9). Le nombre de chromosomes varie d'une espèce à une autre et n'est pas indicatif de la complexité d'un être-vivant. Ainsi, on compte par exemple chez l'Homme 23 paires de chromosomes, 39 chez la Poule, 25 chez l'Ananas, 3 chez le Moustique, 27 chez les Ovins et 30 chez les Bovins et Caprins.

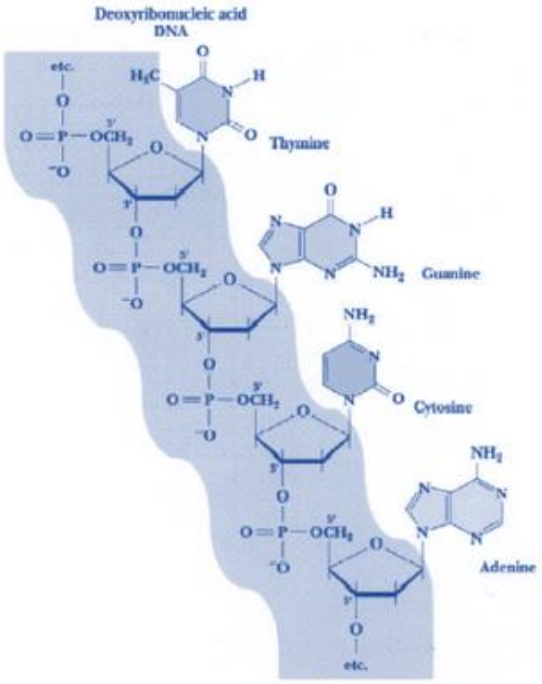


Figure 8: Structure moléculaire de l'ADN
(source : http://www.edu.upmc.fr)

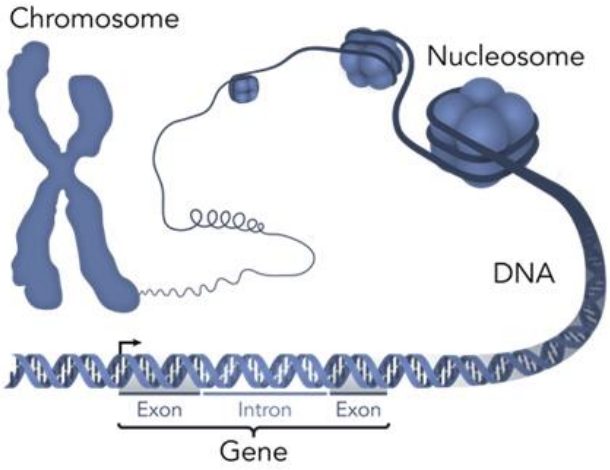


Figure 9: Du chromosome au gène
(source : https://medicalxpress.com/)

NUCLÉOTIDE 1ère POSITION	NUCLÉOTIDE 2ème POSITION U	NUCLÉOTIDE 2ème POSITION C	NUCLÉOTIDE 2ème POSITION A	NUCLÉOTIDE 2ème POSITION G	NUCLÉOTIDE 3ème POSITION
U	UUU phénylalanine UUC phénylalanine UUA leucine UUG leucine	UCU sérine UCC sérine UCA sérine UCG sérine	UAU tyrosine UAC tyrosine UAA non-sens UAG non-sens	UGU cystéine UGC cystéine UGA non-sens UGG tryptophane	U C A G
C	CUU leucine CUC leucine CUA leucine CUG leucine	CCU proline CCC proline CCA proline CCG proline	CAU histidine CAC histidine CAA glutamine CAG glutamine	CGU arginine CGC arginine CGA arginine CGG arginine	U C A G
A	AUU isoleucine AUC isoleucine AUA isoleucine AUG méthionine	ACU thréonine ACC thréonine ACA thréonine ACG thréonine	AAU asparagine AAC asparagine AAA lysine AAG lysine	AGU sérine AGC sérine AGA arginine AGG arginine	U C A G
G	GUU valine GUC valine GUA valine GUG valine	GCU alanine GCC alanine GCA alanine GCG alanine	GAU acide aspartique GAC acide aspartique GAA acide glutamique GAG acide glutamique	GGU glycine GGC glycine GGA glycine GGG glycine	U C A G

Figure 10: Le code génétique

II.2. Les différents types de polymorphismes : origine de la diversité génétique

La mutation est la force évolutive par laquelle se créent les différents types de polymorphismes. Elle se définit comme le processus aléatoire par lequel une nouvelle forme allélique d'un gène apparaît dans une population. Les différentes mutations et leurs fréquences nous renseignent ainsi sur l'histoire évolutive de la population étudiée.

Il existe différents types de polymorphismes :

- Le SNP (Single Nucléotide Polymorphism) est la plupart du temps issu d'une erreur de réplication de l'ADN n'ayant pas été corrigée (fréquence 10^{-7}). Si elle intervient dans une région codante, du fait, de la redondance du code génétique, son impact peut être nul, on parle alors de mutation silencieuse.
- L'insertion/délétion (ou indel) est un ajout ou une suppression de nucléotides (de quelques nucléotides à plusieurs dizaines voire plusieurs centaines de paires de base (pb)). L'impact des indels peut être important car ils modifient le nombre total de nucléotides et ce faisant, décalent complétement le cadre de lecture de la séquence génétique par les ribosomes. Ils peuvent de plus faire disparaître du génome des régions transcrites ou des régions régulatrices de l'expression des gènes.
- Les répétitions sont des variations du nombre de répétitions d'une petite séquence de quelques nucléotides. Dans cette catégorie, on distingue les microsatellites. Ces polymorphismes ont été à l'origine des premières analyses ADN et servent encore dans certains cas particuliers (identification d'un individu à partir de tissus en criminologie par exemple).
- Les Copy Number Variants (CNV) sont des morceaux conséquents d'ADN (taille supérieure à 1 000 paires de bases) dont le nombre de copies successives varie d'un individu à l'autre.
- Les translocations sont des modifications du code génétique par lesquelles un chromosome échange du matériel avec un autre. Certaines translocations sont à l'origine de cancer. C'est le cas par exemple du lymphome de Burkitt dont l'origine est une translocation du gène MYC (Ferry, 2006; D. Liu et al., 2007).
- Les inversions sont une forme de réarrangement du matériel au sein d'un même chromosome. Un segment de ce dernier se retrouve inversé. Les inversions peuvent être paracentriques c'est-à-dire qu'elles n'incluent pas le centromère ou péricentriques lorsque c'est le cas.

Toutes ces variations sont relevées par le séquençage d'un individu et sa comparaison à une séquence de référence (cf III). Les travaux menés au cours de ma thèse se concentrent toutefois sur les plus petites modifications : SNP et petits indels. Les autres types de polymorphismes font, en effet, appel à des techniques particulières d'identification (*calling)*.

II.3. Effets des différents polymorphismes

Lorsque l'effet d'une mutation est largement quantifiable sur le phénotype observé ou entraine une distribution discrète des phénotypes, on parle de gène majeur. Il en existe des exemples dans plusieurs espèces. Certains gènes majeurs liés à l'hyperprolificité des brebis ont été identifiés en ovins (Bindon, 1984; Drouilhet, Lecerf, Bodin, Fabre, & Mulsant, 2009). En caprins, sept haplotypes issus de six polymorphismes dans le gène PRP permettent de définir des profils de résistance à la tremblante. Cette dernière a fait l'objet d'un plan national d'éradication des allèles sensibles (Barillet et al., 2009). Chez le porc, une mutation du gène RYR1 associée à une hyper-musculairité, est à l'origine d'un défaut de la qualité des carcasses appelée PSE (*Pale Soft Exudative syndrome*). Pour un individu polyploïde, le phénotype qui découlera de l'expression de la mutation dépendra du nombre d'allèles mutés qu'il porte et du statut de la mutation (haplo-suffisance, haplo-insuffisance, co-dominance etc…).

De plus, l'effet de la mutation n'est pas le même en fonction de l'endroit où elle se produit. Tout d'abord, d'après le code génétique (Figure 10), une mutation ponctuelle qui aurait lieu en fin de codon n'induit dans la majorité des cas, aucun changement dans l'acide aminé correspondant et ne modifie donc pas la séquence de la protéine exprimée.

La partie codante du génome d'un individu eucaryote est minime. Selon certaines études, cette proportion est même inversement proportionnelle à la complexité d'un organisme. Elle est, ainsi, estimée en moyenne à environ 90% chez les Procaryotes, 68% chez la Levure, 25% pour les Nématodes, 17% chez les Insectes, 9% chez le Poisson-Globe, 2% in la Poule et 1% chez les Mammifères (Taft, Pheasant, & Mattick, 2007). Une mutation qui intervient dans la séquence codante d'un gène peut provoquer un changement d'acide aminé, on parle de mutation faux-sens. Des mutations de ce type ont, par exemple, été repérées en caprin dans le gène CSN1S1 qui code pour la caséine alphaS1, une protéine du lait (Martin and Leroux 2000). La partie codante du gène comporte de nombreux variants (SNPs et insertions/délétions). D'autres variants d'intérêt ont également été identifiés dans le gène DGAT1 qui influence la composition en gras du lait (Martin et al., 2017). Des mutations dans la séquence du gène peuvent aussi induire un arrêt précoce de la traduction ce qui produit une protéine tronquée, on parle de mutation non-sens. L'exemple le plus connu est le groupe sanguin O chez l'Homme qui résulte d'une mutation non-sens dans le gène codant pour l'antigène des hématies.

Il arrive qu'une mutation induise la perte de la fonction initiale de la protéine codée par le gène, on parle de mutation perte-de-fonction. La mutation identifiée dans SOCS2 (Rupp

et al., 2015) est une mutation perte de fonction qui induit une perte d'affinité de la protéine SOCS2 pour son ligand et conduit à une hypersensibilité de l'individu aux mammites. Ces mutations peuvent avoir des conséquences dramatiques. Toutefois, un certain nombre de ces mutations a été retrouvé avec de faibles fréquences chez l'Homme dans le cadre du projet 1 000 génomes. Il a ainsi été estimé qu'un homme est porteur d'une centaine de ces dernières dont 20 à l'état homozygote. Les porteurs ne présentent pour autant pas de maladie particulière. Parmi les pistes envisagées, certains gènes sont en effet en cours de pseudogénisation, c'est-à-dire qu'ils ne codent plus pour une protéine suite à la levée d'une pression sélective. Leur fonction n'est parfois plus capitale au bon fonctionnement de l'organisme. (Monget & Reiner, 2014)

Certaines mutations n'affectent pas directement une protéine mais induisent tout de mêmes des modifications phénotypiques. Ainsi, il existe des régions du génome qui sont transcrites en ARN non-codants. Les micro-ARN (ou miARN) ont par exemple une fonction de régulation de la transcription des ARN messagers d'autres gènes. Ces derniers peuvent avoir des cibles multiples. Ainsi, par exemple, la mutation de l'hypermuscularité en mouton Texel est une substitution qui crée un site d'accueil pour deux miARN (*mir1* et *mir206*) qui modifie l'expression du gène de la myostatine.

Des études chez l'Homme ont montré que 88% des mutations associées à des maladies ou des caractères d'intérêt se trouvent dans des régions introniques ou inter-géniques (Hindorff et al., 2009). Ces régions ne sont donc pas directement liées à une protéine mais induisent tout de même des modifications observables à l'échelle du phénotype. Une mutation si elle intervient dans une zone régulatrice de l'expression d'un ou plusieurs gènes peut avoir des conséquences importantes et mesurables. Ainsi, en bovins, une étude approfondie de la stature des animaux a mis en évidence deux mutations dans une zone intergénique entre PLAG1 et CHCHD7 (Karim et al., 2011). PLAG1 code pour un facteur de transcription qui régule notamment l'expression de facteurs de croissance comme IGF2. Les deux mutations candidates identifiées sont intergéniques et modifient le promoteur des deux gènes (promoteur bi-directionnel).

II.4. Au-delà de la séquence ADN

La séquence d'un individu n'est pas l'unique déterminant des phénotypes. En effet, des interactions complexes avec l'environnement peuvent également modifier l'expression des gènes d'un individu. Ainsi, si un génome correspond à un individu, ce dernier peut avoir plusieurs profils d'expression ou transcriptomes au cours de sa vie et donc plusieurs

protéomes (protéines exprimées à partir des ARN eux-mêmes issus de l'expression des gènes).

L'épigénétique est un mécanisme qui intervient pour modifier l'expression des gènes. Un individu est donc également caractérisé par son épigénome. Cette notion englobe l'ensemble des phénomènes susceptibles de modifier l'expression phénotypique d'un génome sans en altérer sa séquence. Ils sont transmissibles d'une cellule-mère à ses cellules-filles et également potentiellement transmissibles à la descendance d'un individu (épigénétique transgénérationnelle). La définition générale de l'épigénétique est donc relativement large. Toutefois les mécanismes de cette modification de l'expression sont identifiés : il s'agit de la méthylation de l'ADN, la modification des histones (méthylation, acétylation) et l'intervention d'ARN non-codants. Ces modifications peuvent être réversibles. (Goldberg, Allis, & Bernstein, 2007; Monget & Reiner, 2014). Certains phénotypes particuliers trouvent leur origine dans des modifications épigénétiques qui aboutissent à des modifications de la conformation de la chromatine (et donc de l'expression des gènes). Les épi-mutants sont rares, toutefois, quelques-unes de ces épimutations sont connues. Par exemple, chez la Tomate (*Solanum lycopersicum*) une hyperméthylation de la région promotrice du gène LeSPL-CNR induit une perte de maturation du fruit (absence de coloration, problème d'adhésion cellulaire) (Chen et al., 2015). Cette mutation fait toutefois partie des épiallèles facilités c'est-à-dire dépendants d'un génotype particulier.

Les pré-ARN peuvent être modifiés une fois transcrits selon plusieurs mécanismes (ajout d'une coiffe, ajout d'une queue poly-A, épissage etc…) qui ont pour but de protéger l'ARN ou d'en modifier la séquence. L'épissage modifie la chaine en excisant des morceaux d'ARN (les introns) de la séquence ou en suturant deux ARN. Un épissage alternatif du même gène peut donc conduire à deux protéines de structures et fonctions différentes répondant à des besoins différents. Ainsi, il a été prouvé que le VIH (virus de l'Immunodéficience Humaine) est capable de synthétiser une multitude d'ARN matures différents à partir du même pré-ARN. Les protéines qui en découlent peuvent impacter la multiplication ou l'infectiosité du virus (Purcell & Martin, 1993).

Enfin certaines modifications protéiques ont lieu après la polymérisation de la chaine d'acides aminés. La conformation de la protéine peut alors se trouver altérée sans que la séquence ADN associée n'ait été modifiée. Ces modifications sont la plupart du temps assurées par des protéines. On peut citer plusieurs mécanismes comme le clivage de chaîne polypeptidique, la formation de ponts covalents, l'ancrage lipidique, la glycosylation, la

phosphorylation etc… L'insuline fait partie de ces protéines qui subissent des modifications post-traductionnelles (Figure 11). Le clivage de la protéine a été mis en évidence lors de la découverte de la voie de biosynthèse de l'insuline (Steiner, Cunningham, Spigelman, & Aten, 1967). L'expression du gène de l'insuline conduit à la synthèse de pro-insuline. Une fois repliée, cette dernière est clivée pour former l'insuline. Le clivage produit un peptide qui est notamment utilisé en médecine pour le suivi de certaines formes de résistance à l'insuline.

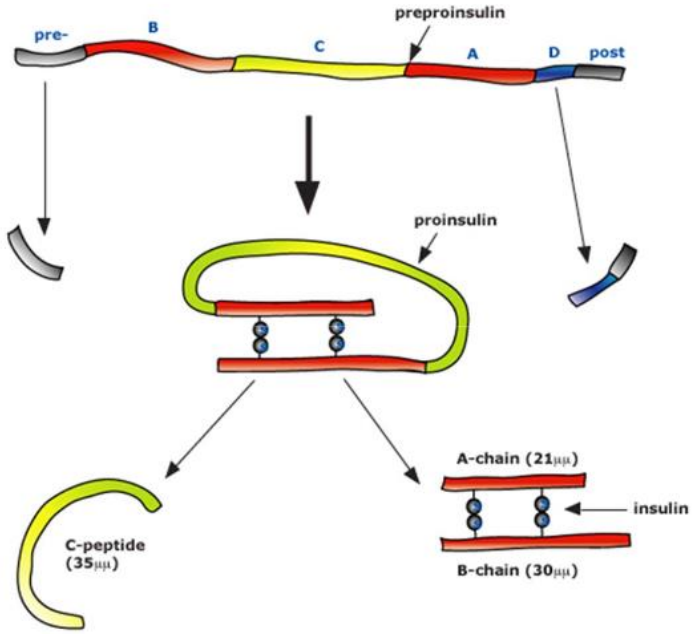


Figure 11: Exemple de modification post-traductionnelle : l'insuline
(source: unf3s.cerimes.fr)

III. Le séquençage

III.1. De la méthode de Sanger au Next Generation Sequencing

La méthode Sanger est une des premières méthodes développées ayant permis d'accéder à la séquence de nucléotides d'un individu. Elle repose sur l'incorporation un à un des nucléotides à partir d'une amorce. En plus des nucléotides « normaux », on fournit au complexe de réplication de l'ADN des nucléotides modifiés appelés ddNTPs, le groupement OH du désoxyribose est remplacé par un H. A chacune des quatre bases est associé un fluorochrome différent. A chaque fois qu'un de ces nucléotides est ajouté à la séquence, il termine la chaîne de polymérisation de l'ADN. L'incorporation des bases se fait au hasard parmi un pool de nucléotides marqués ou non. L'arrêt de la synthèse est donc aléatoire et permet d'obtenir des brins d'ADN de tailles diverses. En fin de réaction, les différents

fragments produits sont soumis à une électrophorèse qui révèle alors la succession des différentes bases (Figure 12).

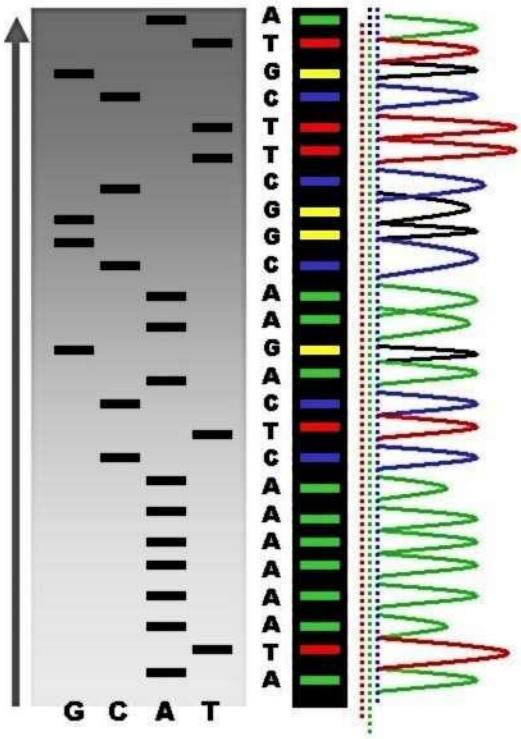


Figure 12: Exemple de résultat d'un séquençage de Sanger
(source : futura-sciences)

Cette méthode a bénéficié du développement de nouvelles technologies qui en a accru les performances : utilisation de la PCR pour l'amplification des fragments, séquenceurs automatiques pour lire la succession des bases. Avec cette première approche, les fragments mesuraient en moyenne 100 à 400 pb et pouvaient même atteindre 1 000 pb. Cette technique a permis d'assembler le premier génome : celui du Bacteriophage Phi-X174 en 1977.

Bien que la méthode de Sanger ait prouvé sa fiabilité, elle ne permet encore d'analyser qu'un faible nombre de séquences de petite taille et est extrêmement chronophage. Elle n'est donc plus utilisée que pour le séquençage ciblé d'un fragment d'ADN précis. Plus récemment, de nombreuses méthodes ont vu le jour pour améliorer l'efficacité du séquençage et en réduire le coût.

III.2. Le Next Generation Sequencing

Les premiers séquenceurs Next-Generation Sequencing (NGS) sont apparus en 2007 et avec eux, le coût de séquençage a fortement chuté (Figure 13). Le NGS est une technique qui repose sur le morcellement aléatoire de l'ADN et sur une grande capacité de parallélisation de la lecture de la séquence ADN.

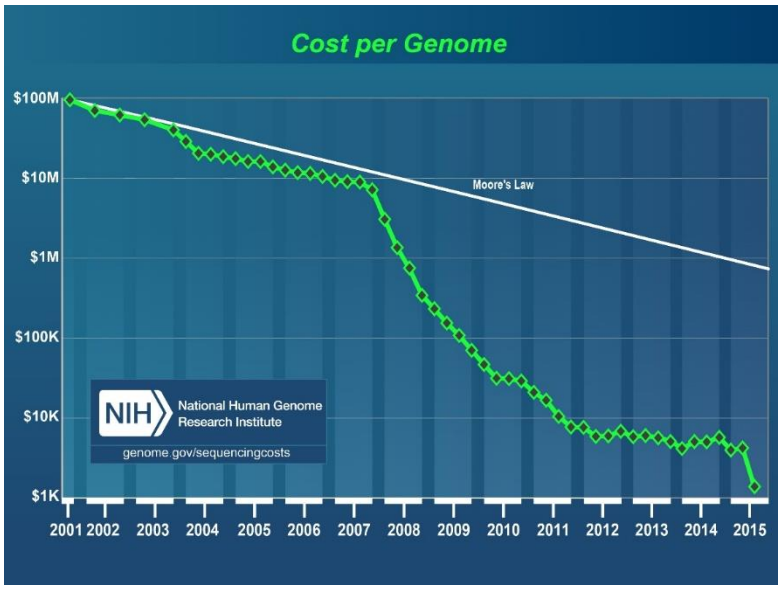


Figure 13: Evolution du coût de séquençage avec l'apparition des séquenceurs haut-débit
(source : National Human Genome Research Institute)

III.2.a. Le fonctionnement des séquenceurs haut-débit

La force des séquenceurs haut-débit est qu'ils sont capables de traiter en parallèle plusieurs fragments d'ADN de quelques dizaines à quelques centaines de paires de bases. Ces fragments sont appelés lectures ou *reads*. A chacun des fragments est associée une paire de séquences adaptatrices. Elles permettent la fixation de ces derniers à un support solide. Chaque fragment est ensuite amplifié. Les amplificats résultants (environ 1 000) sont séquencés en parallèle. A chaque fois qu'une base est ajoutée à la chaine de réaction, un signal fluorescent spécifique du nucléotide ajouté est émis et lu par le séquenceur. La nature de la réaction dépend de la technologie de séquençage utilisée. (Liu et al., 2012)

III.2.b. Traitement bioinformatique des données en sortie du séquenceur

En sortie de séquençage nous obtenons des fichiers traités pour obtenir des *fastq* (Cock, Fields, Goto, Heuer, & Rice, 2010). Les *fastq* sont des fichiers compacts mais lourds car ils contiennent de multiples informations sur les lectures obtenues par le séquenceur. Le fichier contient 3 lignes par lecture : La première ligne commence systématiquement pas un

« @ » et contient l'identifiant de l'échantillon séquencé, vient ensuite une ligne qui explicite la séquence de nucléotides, puis une ligne contenant le symbole « + » et enfin une ligne comportant une chaine dont chacun des caractères se rapporte à une des bases lues et nous informe de sa qualité. Cette qualité est un PHRED score (Q_{PHRED}) qui se définit selon la formule :

$$Q_{PHRED} = -10 \times log_{10}(P_e)$$

Où P_e représente la probabilité d'avoir une erreur de séquençage. Chaque base qui apparaît dans la lecture est une estimation de la base véritable, cette estimation est appelée *calling* et est soumise à des erreurs. La probabilité d'une erreur est estimée par le logiciel PHRED à la lecture des fichiers de séquençage (Ewing et al., 1998a; Ewing et al., 1998b). Le logiciel estime alors la qualité du « *calling* » c'est-à-dire la probabilité que le bon nucléotide ait été lu. Ainsi une probabilité de 1% d'erreur correspond à un Q_{PHRED} de 20.

Les fichiers *fastq* nourrissent ensuite un programme d'alignement des lectures. Ce logiciel lit les séquences des lectures et cherche à les aligner sur un génome de référence. Cet alignement peut ne pas être parfait, c'est-à-dire qu'on autorise des non-concordances et des « trous » qui peuvent correspondre à des insertions/délétions ou à des SNP. Un séquenceur Illumina et Solexa produit en moyenne 50 à 200 millions de lectures dont la longueur est comprise entre 32 et 100 paires de bases (Li & Durbin, 2009). La répartition de ces lectures sur le génome est aléatoire. L'alignement de chacune est exigeant du point de vue computationnel car il faut tester un très grand nombre de possibilités. Le logiciel d'alignement produit un fichier *SAM* (Sequence Alignment/Map). Ce fichier contient l'information d'alignement pour chacune des lectures : la séquence du read, sa position d'alignement optimale sur la séquence de référence, la qualité de cet alignement, la présence ou non de variations par rapport à la référence ainsi que leur type (insertion, délétion, variation d'un nucléotide etc…). La qualité d'alignement ou MQ pour Mapping Quality se définit suivant l'équation suivante :

$$MQ = -10 \times log_{10}(P_m)$$

Où P_m représente la probabilité que l'alignement soit incorrect.

Les fichiers *SAM* donnent des informations sur les lectures les unes par rapport aux autres. Ainsi pour une lecture on trouve également la position relative de la lecture suivante et

son identifiant. Des *flags* sont aussi associés à chaque lecture pour donner des informations par rapport à sa paire (autre lecture) si elle existe. On peut ainsi savoir si une paire existe et si elle est correctement alignée ou non. Ces fichiers sont compressés sous un format binaire appelé *BAM*.

Les fichiers *BAM* sont un support pour les programme de *calling* : GATK (McKenna et al., 2010), SAMtools mpileup (Li et al., 2009), FreeBayes (Garrison & Marth, 2012) par exemple. Avant de faire le *calling* définitif des variants, il faut toutefois enlever les duplicats des jeux de données c'est-à-dire les lectures identiques qui peuvent facilement propager une erreur (Figure 14).

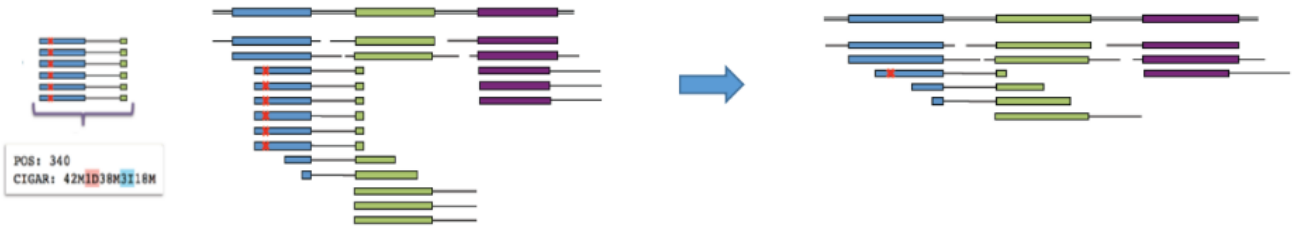


Figure 14: Elimination des duplicats dans les lectures avant le calling définitif des variants
(source : https://ming-lian.github.io/2019/02/08/call-snp/)

Ensuite un réalignement local est effectué pour minimiser le nombre de bases sur l'ensemble des lectures qui ne correspondent pas à celles de la référence (bien qu'alignées). Enfin une recalibration des bases ou BQRS (Base Quality Recalibration Score) est appliquée. Cette étape est recommandée par le Broad Institute (GATK best practices). En effet, bien que les séquenceurs attribuent un score de qualité relativement fiable à chaque base lue (Phred score des *fastq*), ils sont aussi sujets à des erreurs systématiques (défaut de la machine ou problème lié à la réaction chimique du séquençage). Ces erreurs systématiques sont modélisées à l'aide de technologie d'Intelligence Artificielle telle que le machine learning. Le machine learning est fait sur un set de variants connus. Il permet *in fine* de calculer une qualité de *calling* plus précise. Le *calling* en lui-même ne peut pas être modifié à cette étape, toutefois, la réévaluation de sa qualité permet de tenir compte d'une potentielle erreur lors de l'agrégation des fichiers *BAM*.

Après toutes ces étapes, un fichier *BAM* final est créé. Les informations qu'il contient permettront de former un *gvcf* (variant calling format). Le *gvcf* résume toutes les informations

de la séquence d'un individu et son génotype à toutes les positions du génome qu'il soit différent de la référence ou non et qu'il y ait une lecture alignée ou non (le cas échéant il y aura un génotype manquant et aucun score de qualité). Ce fichier est un fichier compressé individuel. Un logiciel de *calling* (GATK (McKenna et al., 2010), SAMtools mpileup (Li et al., 2009), FreeBayes (Garrison & Marth, 2012) etc…) tient compte de toutes les informations présentes dans chaque fichier *BAM* pour identifier un variant (Tableau 4). La notion de profondeur est importante à cette étape. La profondeur peut être définie comme le nombre de lectures qui s'alignent à un endroit donné du génome. Ainsi, si 10 lectures s'alignent à un endroit, la profondeur sera de 10. Le génotype d'un individu à cette position sera déduit du nombre d'observations des différents allèles à la position. Par exemple, avec une profondeur de 10, si la séquence d'un individu contient 9 lectures de l'allèle de référence (REF) et 1 de l'allèle alternatif (ALT), le logiciel de *calling* définira l'individu comme homozygote REF à la position.

Tableau 4: Informations utilisées par le logiciel de calling et l'échelle à laquelle elle se rapporte

ECHELLE DE L'INFORMATION	***INFORMATION***
LA BASE NUCLEOTIDIQUE	Phred score recalibré
LA LECTURE	Mapping quality
	Brin d'ADN concerné (forward/reverse)
LA POSITION	Allèles de la position (REF pour la référence, ALT pour un allèle alternatif à ce dernier)
	Nombre d'observations de ALT et REF
	Profondeur de séquence
LE GENOTYPE	Génotype d'un individu déduit du ratio (REF/ALT)

L'ensemble des *gvcf* d'individus analysés conjointement peut être agrégé par un logiciel de *calling* sous la forme d'un fichier *vcf* qui lui ne contient plus que ces informations pour les positions du génome pour lesquels des allèles alternatifs à la séquence de référence ont pu être observés (au minimum un génotype hétérozygote sur l'ensemble des individus analysés). Ainsi, on y trouve des informations générales sur le variant : chromosome, position sur le génome, allèles observés. Une qualité de la position (QUAL) est également estimée, elle représente la probabilité qu'on ait bien observé un polymorphisme à cette position. Ainsi, une qualité de 100 indique probabilité d'erreur de 1 sur 10^{10}. Ce paramètre ne peut pas être utilisé comme seule critère de filtrage car il atteint rapidement des valeurs élevées lorsque le nombre d'individus intégrés dans le *vcf* est grand. On trouve également pour chaque variant

des informations sur les individus séquencés : génotype, profondeur de séquence à la position, nombre de lectures pour chacun des allèles, la qualité du génotype (PHRED score incluant la probabilité d'une erreur estimée par le logiciel de *calling*).

III.2.c. Construction d'une séquence de référence

Les traitements que nous avons présentés précédemment reposent pour partie sur l'alignement des lectures sur une séquence de référence. En l'absence d'une séquence de référence, il est possible de reconstruire une séquence *de novo* à partir des données de lectures fournies par un séquenceur. Pour maximiser la couverture obtenue sur le génome de l'individu de référence (c'est-à-dire la profondeur moyenne de séquence), il est préférable de fournir suffisamment de matériel ADN. L'assemblage demande, en effet, une profondeur locale plus importante qu'un séquençage à des fins de génotypage et de comparaison à un assemblage existant pour avoir en moyenne une profondeur de 30X. Elles peuvent être plus ou moins longues et sont issues de la fragmentation aléatoire du génome d'un individu. L'assemblage de toutes ces lectures pour former une seule séquence est assurée par un algorithme. Il fonctionne selon le principe suivant :

- Il calcule toutes les possibilités de combinaisons 2 à 2 des lectures.
- Il choisit la paire de lectures pour laquelle le chevauchement est le plus grand.
- Il fusionne ces deux lectures pour obtenir un fragment plus long qui sera par la suite considéré comme une seule lecture.
- Il répète les deux étapes précédentes jusqu'à ce qu'il n'ait plus la possibilité de former une paire.

Un ensemble de lectures assemblées est appelé *contig*. Il n'est pas possible de relier directement les *contigs* qui sont, par définition, non-chevauchants. On utilise alors de longs fragments d'ADN dont on ne séquence que les extrémités. Ces dernières peuvent alors permettre d'assembler deux contigs (Figure 15). La longueur de ces fragments étant connue, ils permettent de plus de construire une carte physique de l'espèce étudiée. Un ensemble de *contigs* est appelé *scaffold*. Les *scaffolds* peuvent ainsi être formés sur le même principe que les *contigs* à l'exception près qu'un *scaffold* peut contenir des « trous » (Figure 15). Les *scaffolds* peuvent éventuellement former des chromosomes.

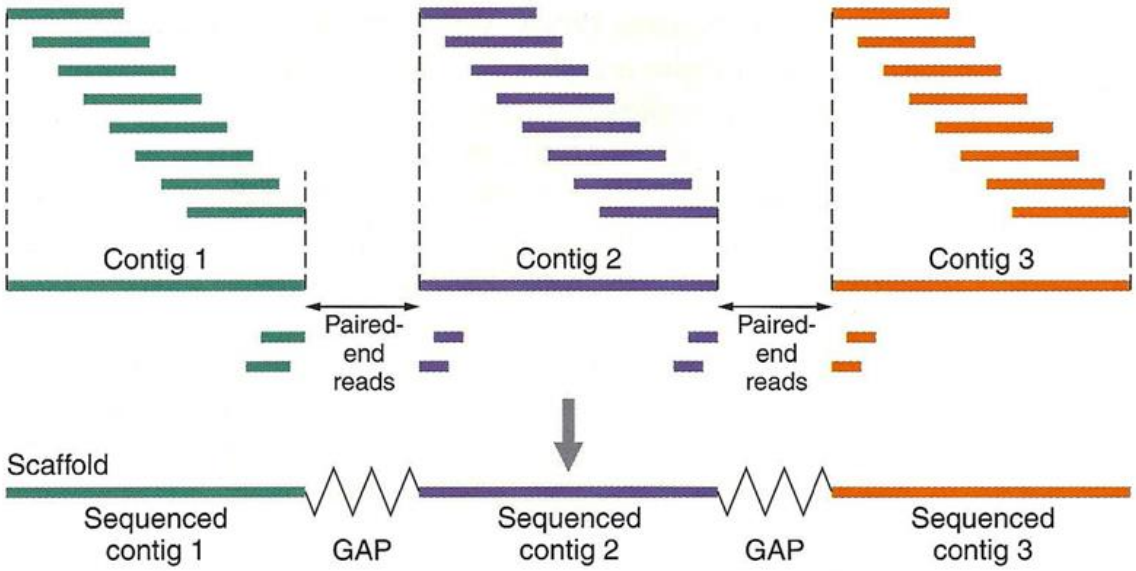


Figure 15: Formation des contigs et scaffolds à partir de données NGS
(source : discoveryandinnovation.com)

III.3. Les promesses du séquençage

Le séquençage d'un nombre limité d'individus choisis pour représenter au mieux la population d'une espèce a précédemment permis d'identifier des zones polymorphes d'intérêt qui ont conduit à l'élaboration d'outils de génotypages. De tels outils ont été développés dans nombre d'espèces et permettent d'obtenir de manière fiable une information ponctuelle sur le génome d'un individu pour un coût raisonnable. Les microsatellites qui, rappelons-le sont des répétitions de courtes séquences dont le nombre de répétitions est variable d'un individu à un autre, ont été les premiers polymorphismes à être utilisés pour le développement de tels outils. Par la suite, ce sont des marqueurs liés à des polymorphismes de type SNP qui ont été utilisés pour développer les puces à ADN. Ces derniers sont largement plus répandus sur le génome que les microsatellites mais, contrairement aux microsatellites, ne présentent pas plus de 2 allèles. Les puces construites sur la base des SNP peuvent avoir diverses densités : de quelques centaines (puces d'assignation de parenté par exemple) à plusieurs centaines de milliers. En caprins, actuellement, seule une puce de 53 347 marqueurs est disponible pour l'espèce (G Tosser-Klopp et al., 2012).

Une puce ADN est une sélection de SNP choisis pour balayer uniformément le génome. Elle est, par définition, incomplète. Le séquençage, quant à lui, permet en théorie d'accéder à l'intégralité de la variabilité génétique structurelle d'un individu. Ainsi, comme nous l'avons souligné précédemment, en plus des SNP qui sont un changement ponctuel de nucléotide, d'autres types de variants peuvent être identifiés sur la séquence : les

insertions/délétions (ou indel) qui sont des modifications plus ou moins étendues et les CNV (Copy Number Variants) qui correspondent à des motifs répétés de longueur variable.

Aussi, si suffisamment d'individus sont séquencés et phénotypés pour un caractère donné, une analyse d'association peut potentiellement révéler le variant causal d'un phénotype donné. Toutefois, dans les filières animales comme ailleurs, le coût du séquençage n'a encore pas suffisamment diminué pour qu'un très grand nombre d'individus soient séquencés en routine. Une façon efficace d'utiliser les séquences disponibles est de recourir à l'imputation. L'imputation permet de prédire statistiquement la séquence d'individus qui ont été génotypés à l'aide de puces ADN, plus accessibles d'un point vue économique, à partir des informations d'animaux de la même population qui sont eux séquencés.

Dans la majorité des filières et pour une grande partie des caractères, à défaut d'avoir pu identifier une mutation causale pour un phénotype particulier, la sélection génomique repose sur le déséquilibre de liaison entre le variant causal et le marqueur retenu sur le panel. Il est donc statistiquement possible qu'une recombinaison ait lieu entre le marqueur et la mutation causale ce qui briserait localement le déséquilibre de liaison. La probabilité d'un tel évènement dépend de la proximité du marqueur avec la mutation causale mais aussi de l'historique de la race (effets de dérive et sélection). Par conséquent, si le variant causal est identifié, il peut être utilisé en sélection et en améliorer la précision. La précision de la prédiction génomique serait alors conservée sur plusieurs générations.

III.4. Les limites au séquençage

Les séquenceurs commettent très peu d'erreurs dans la lecture de la succession des bases notamment en *short reads*, en effet, les dernières générations sont fiables à 99,9% (Liu et al., 2012; Quail et al., 2012). Toutefois, si ce pourcentage est mis en perspective avec les 3 Gbp du génome caprin, il y a potentiellement 3 millions d'erreurs de lecture par individu séquencé.

L'acquisition de données de séquence pour un individu repose sur la capacité du logiciel de *calling* à aligner correctement les lectures sur la séquence de référence. Les régions contenant des motifs répétés ou particulièrement enrichies en une base nucléotidique donnée peuvent être problématiques. De telles régions sont représentées à divers endroits du génome. Il y a débat dans la communauté scientifique quant à la proportion de séquences répétées (qu'elles soient complexes comme les transposons ou simples comme dans les *tandem repeats*) dans le génome. Chez l'Homme elle a pu être estimée comme couvrant près de 2/3

du génome (Koning, Gu, Castoe, Batzer, & Pollock, 2011). Les motifs répétés rendent la création d'une séquence de référence ou l'alignement d'une séquence compliquée et incertaine. En effet, une lecture couvrant un morceau de séquence répétée peut s'aligner à différents endroits du génome. Les mutations qu'elle pourrait révéler ne seraient pas positionnées avec certitude au bon endroit du génome.

Enfin la profondeur de séquençage du génome est un paramètre important pour obtenir une bonne estimation des génotypes d'un individu. Plus elle est importante, plus le génotype obtenu à une position est fiable. En effet, le poids d'une lecture est inversement proportionnel au nombre de lectures qui s'alignent à une position équivalente. Prenons un individu homozygote à un locus donné, si une erreur de séquençage se glisse sur une de ses lectures, si seulement 2 lectures s'alignent à une position donnée alors le logiciel de calling aura tendance à lire un hétérozygote. Si la séquence avait été localement plus profonde alors le logiciel de *calling* aurait pu détecter qu'une des lectures était porteuse d'une erreur et aurait trouvé le véritable génotype de l'individu. Ainsi, lorsque peu de lectures sont alignées sur une position, il est plus difficile de savoir si l'on a bien détecté un allèle alternatif ou si le polymorphisme observé est lié à un mauvais alignement ou une erreur de séquençage d'une des lectures.

Enfin pour détecter les variants rares dans une population il faut qu'au minimum un individu dans l'analyse soit porteur de l'allèle rare. Ainsi pour espérer détecter un variant dont la fréquence de l'allèle minoritaire (ou MAF pour *Minor Allele Frequency*) est de 1%, il faut séquencer au moins 50 individus pour détecter un hétérozygote.

III.5. L'annotation du génome

L'annotation consiste, une fois le génome séquencé, à chercher les éléments d'intérêt : gènes, zone régulatrices etc… Il s'agit d'une étape cruciale pour les analyses ultérieures, car elle peut permettre de comprendre le lien entre une mutation candidate et le phénotype étudié. Elle permet aussi de prioriser différents variants causaux possibles en fonction de leur nature (intron, exon, zones intergéniques...) et de l'activité de la protéine associée au gène, le cas échéant. En caprins, ce sont 21 361 gènes codants qui ont été identifiés sur le génome (www.ensembl.org/Capra_hircus/Info/Annotation; consulté le 10/04/2020).

L'annotation est spécifique d'une espèce, en effet, les positions des gènes ne sont pas identiques inter-espèces. Toutefois, en l'absence d'annotation exhaustive comme c'est le cas en caprin, il est possible de chercher des gènes connus sur des espèces voisines dont le génome est annoté. C'est une recherche basée sur l'homologie de séquence entre les gènes

d'espèces distinctes. D'autres méthodes sont disponibles et s'appuient sur la structure même d'un gène qui commence systématiquement par un codon-start ATG, marquant le début d'une zone transcrite, et s'achève par un codon-stop TAA, TAG ou TGA qui induit le décrochement des ribosomes responsables de la transcription de l'ADN en ARN. En pratique, les codons-start et codons-stop sont automatiquement détectés dans la séquence. Des algorithmes signalent alors la présence supposée d'un gène lorsque l'un ou l'autre des codons est observé. Lorsque le gène ainsi prédit est inconnu, un identifiant lui est attribué sans qu'aucune fonction ou protéine ne lui soit encore associée.

Deux projets internationaux ont pour but de caractériser les éléments fonctionnels des génomes assemblés. Ils utilisent plusieurs techniques de pointe pour arriver à leurs fins : le RNA-seq qui permet d'avoir une idée de l'expression de zones spécifiques du génome, le Hi-C et l'ATAC-seq qui nous informent sur la proximité spatiale des chromosomes. Le projet ENCODE d'annotation du génome humain implique plusieurs centaines de chercheurs des Etats-Unis, du Royaume-Uni, d'Espagne, de Singapour et du Japon. Le projet FAANG (Functional Analysis of Animal Genomes) est son pendant international chez les animaux d'élevages. Il a pour but de coordonner les efforts pour faciliter la recherche d'un lien entre génotype et phénotype. L'application du projet FAANG en France est le projet Fr-AgENCODE, financé par le metaprogramme SELGEN INRAE (Foissac et al., n.d.). Quatre espèces sont concernées : le porc, la poule, la vache et la chèvre. L'annotation du génome caprin est également soutenue par l'IGGC dans le cadre de plusieurs groupes de travail du projet VarGoats.

IV. L'imputation

L'acquisition de données de séquence restent à l'heure actuelle coûteuse. Une façon d'exploiter au mieux la petite population séquencer est de recourir à l'imputation. Cette méthode se définit comme la capacité de prédire les génotypes manquants dans une population à partir des génotypes observés dans cette population et dans une population de référence génotypée à plus haute densité. L'imputation se définit comme une méthode par laquelle on prédit les génotypes manquants dans une population à partir des génotypes déjà observés dans cette même population et des génotypes d'une population de référence typée à plus haute densité (Boichard et al., 2012).

Bien qu'aucune étude préalable à cette thèse n'ait été conduite sur l'imputation vers la séquence en caprins, cette thématique a largement été abordée dans d'autres filières. On trouve ainsi des études d'imputation vers la séquence en volaille (Heidaritabar, Calus,

Vereijken, Groenen, & Bastiaansen, 2015; Ye et al., 2018), bovin laitier (Frischknecht et al., 2017; Pausch et al., 2017; Sahana, Guldbrandtsen, Lund, & Vilkki, 2016; Van Binsbergen et al., 2014) et allaitant (Daetwyler et al., 2014; Li et al., 2014), en ovin (B. J. Hayes, Bowman, Daetwyler, Kijas, & Van Der Werf, 2012) ainsi qu'en porc (van Son et al., 2017; Zhang et al., 2018).

IV.1. Principe général

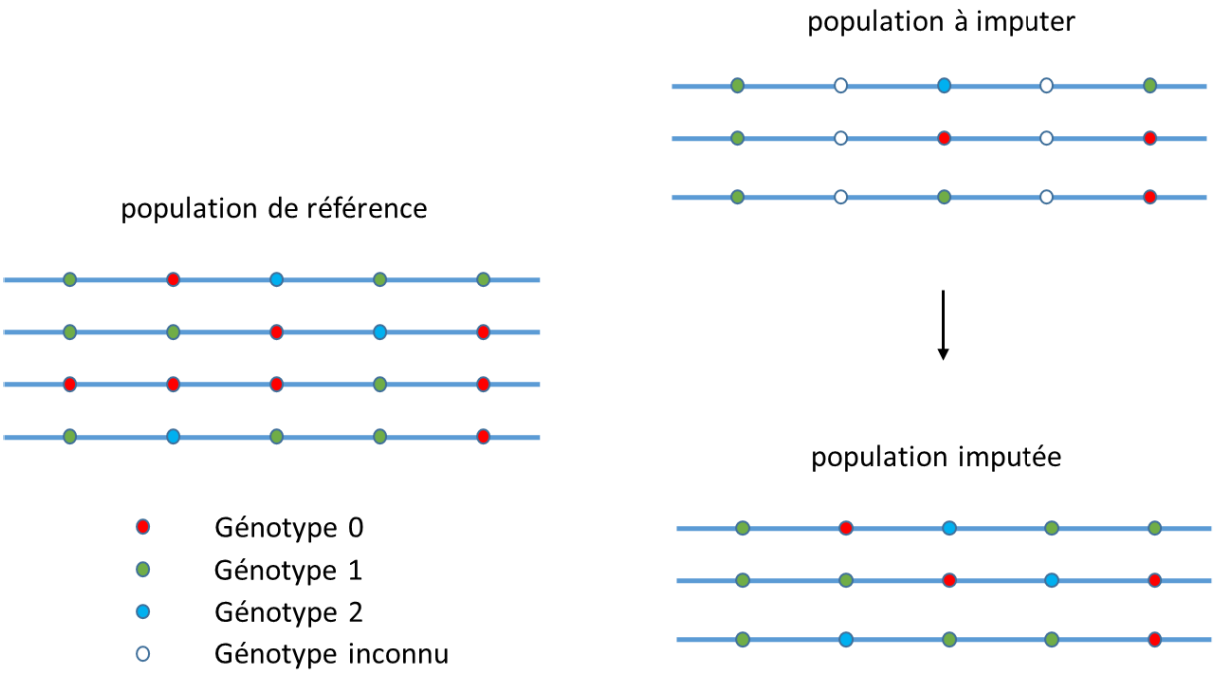


Figure 16: Principe de la prédiction des génotypes inconnus par imputation

Le principe de l'imputation, schématisé la Figure 16, consiste à combler l'information manquante (positions génomiques non génotypées) avec des génotypes les plus probables (best-guess genotypes) ou des probabilités de génotype (genotype dosages) en s'appuyant sur le déséquilibre de liaison entre marqueurs successifs. La population de référence est utilisée pour construire une bibliothèque d'haplotypes, c'est-à-dire de génotypes (allèles de différents marqueurs) fréquemment associés, et calculer leur fréquence. Les génotypes connus des typages à imputer sont ensuite comparés à la bibliothèque d'haplotypes et l'haplotype le plus probable est sélectionné pour un individu donné. Une fenêtre glissante parcourt ainsi le typage de chaque individu, compare ce qu'elle observe à la bibliothèque, prédit les génotypes et met à jour les fréquences des haplotypes dans la bibliothèque à mesure qu'elle avance. Lorsque le pedigree est disponible, il est possible d'utiliser également l'information d'apparentement des individus pour imputer leurs génotypes à partir de leurs apparentés génotypés en haute densité.

IV.2. Intérêts de recourir à l'imputation

Il est fréquent que le traitement des échantillons ADN ne permette pas d'identifier le génotype de l'intégralité des SNP du panel utilisé pour un individu (problème de qualité de l'échantillon…). L'imputation est alors utile pour compléter les quelques génotypes manquants qui peuvent apparaître en sortie de laboratoire. Elle améliore alors la qualité des typages et, parfois, corrige les éventuelles erreurs de génotypages ou incohérence mendéliennes.

Cette technique est utilisée en routine dans certaines filières comme en bovins laitiers chez qui elle permet notamment d'harmoniser les typages parmi une grande variété de puces disponibles. Les puces de différentes densités d'une même espèce sont conçues pour que les marqueurs des puces de plus basse densité soient également inclus sur les puces de plus haute densité. Tous les génotypes disponibles peuvent alors être imputés vers une même densité de marqueurs. Ils sont utilisés comme une seule et même population pour les analyses ultérieures (évaluations génomiques notamment).

D'un point de vue économique, l'imputation permet de réduire les coûts d'acquisition de typages haute densité. Il est alors possible de génotyper un grand nombre d'individus en basse densité puis de les imputer pour obtenir une plus large population d'individus possédant un génotype haute densité. Elle nécessite toutefois l'acquisition préalable d'un certain nombre de typages haute densité d'individus soigneusement choisis pour constituer la population de référence.

IV.3. Evaluation de la qualité d'imputation

Le processus d'imputation n'est pas infaillible et repose principalement sur des probabilités d'association entre les marqueurs à imputer et les marqueurs observés. Il est, par conséquent, nécessaire d'évaluer la qualité des typages produits par imputation. Cette évaluation se fait en plusieurs étapes génériques. Tout d'abord, les génotypes d'un ou plusieurs génotypages haute densité sont masqués jusqu'à obtenir une densité en marqueurs équivalente à celle de la population à imputer. Puis, ces génotypes sont imputés *de novo* grâce au reste de la population de référence. Il suffit alors de comparer les génotypes imputés aux génotypes vrais des marqueurs dont les génotypes ont été précédemment masqués. La comparaison est traditionnellement effectuée en calculant des corrélations entre génotypes imputés et génotypes vrais ou en calculant des taux d'erreurs d'imputation alléliques ou génotypiques. Ces métriques de qualité d'imputation sont calculées à l'échelle de l'individu ou du SNP.

IV.4. Facteurs influençant la qualité d'imputation

De nombreux travaux se sont attachés à étudier différents facteurs pouvant influencer la qualité d'imputation (Bolormaa et al., 2019; Bouwman & Veerkamp, 2014; Brøndum et al., 2014; Dassonneville et al., 2012; Frischknecht et al., 2017; Hayes et al., 2012; Hozé et al., 2013; Van Binsbergen et al., 2014; Ventura et al., 2014, 2016; Wolc et al., 2011).

IV.4.a. Taille et composition du panel de référence

Comme souligné précédemment, un panel de référence avec une densité en marqueurs élevée est nécessaire pour imputer une population de densité plus faible. Le nombre d'individus dans ce panel de référence est un paramètre important pour l'imputation. Plus la taille de ce dernier est importante plus elle peut être représentative de la structure de la population à imputer et plus l'algorithme d'imputation calcule des fréquences d'haplotypes précises et impute les génotypes manquants avec précision. Van Binsbergen et al. (2014) ont imputé des séquences de bovins de race Holstein avec 3 populations de référence de tailles différentes (respectivement 40%, 60% et 80% des 114 Holstein séquencées disponibles) et ont montré que la qualité d'imputation était accrue avec un nombre d'individus séquencés croissant. Il est à noter toutefois que la qualité atteint rapidement un plateau. Toutefois, pour des variants avec de petites MAF (< 5%), il faut un très grand panel de référence pour arriver à saisir ce plateau. Dans le cas de l'étude précédente, le passage d'une population comportant 60% des animaux disponibles à 80% ne permet pas un gain substantiel de la qualité de séquence imputée. Li et al. (2014) ont étudié l'imputation vers la séquence dans plusieurs races bovines. Certaines d'entre elles avaient des effectifs comparables aux effectifs séquencés en Alpine et Saanen françaises : 25 séquences en race Limousine, 27 en Jersey ou 43 en Brown Swiss. Ils ont obtenu des taux de concordance entre génotypes vrais et génotypes imputés compris entre 75,3% et 84,5% et des corrélations (R^2) comprises entre 0,63 et 0,76 pour une imputation de la 54k vers la séquence. Ces résultats nous laissent envisager qu'une telle imputation est possible en Alpine et Saanen.

La composition de la population de référence est également un paramètre stratégique de la qualité d'imputation. En effet, la présence d'individus représentatifs de la même race ou apparentés à la population à imputer augmente la probabilité de capturer un haplotype présent dans la race ou dans la famille d'un individu à imputer. L'apparentement induit également la conservation de grands haplotypes d'une génération à la suivante ce qui rend plus facile l'imputation. Dans le cas où le nombre d'individus typés en haute densité serait restreint pour une race, il a été montré que l'ajout d'autres races au panel de référence peut améliorer significativement la qualité d'imputation, à condition que le DL entre races soit conservé.

Cependant, si les races s'avèrent génétiquement éloignées, la qualité d'imputation peut aussi s'en trouver dégradée (Frischknecht et al., 2017). Druet et al. (2014) ont étudié plusieurs stratégies de choix des individus à séquencer pour améliorer la qualité d'imputation vers la séquence. Ainsi 3 grandes stratégies se détachent d'un choix aléatoire : (i) maximiser le nombre de génomes indépendants (ii) maximiser l'apparentement entre populations de référence (séquencée) et population à imputer et enfin (iii) maximiser la représentation des haplotypes de la population. Ces 3 stratégies ont été conjuguées pour choisir les animaux séquencés dans le jeu de données VarGoats.

IV.4.b. Densités de typage et déséquilibre de liaison

La différence de densité entre le génotype de départ et le génotype attendu en sortie d'imputation joue un rôle important. La population de référence est potentiellement à adapter en fonction de cette différence de densité. Ainsi, plus la densité de départ est faible, plus il y a de recombinaisons possibles et moins le DL est conservé. Il faut alors une population de référence proche (par exemple la génération antérieure y compris les parents). Si la densité de départ est déjà forte, il y a peu de recombinants, la population de référence peut alors être distante car le DL est conservé. Plus généralement, il est difficile d'imputer avec certitude un génotype manquant à partir d'un génotype connu éloigné de ce dernier. En effet, l'imputation repose sur le phasage des données de génotypages, c'est-à-dire la reconstruction d'haplotypes qui sont des morceaux de chromosomes parentaux qui contiennent les allèles aux marqueurs transmis ensemble. Le phasage est performant de proche en proche car il y a peu de chance d'avoir une recombinaison entre deux marqueurs proches. Ainsi, il est facile de reconstruire des combinaisons d'allèles autour des marqueurs de la puce. Dès que l'on s'éloigne d'un marqueur, l'incertitude sur le phasage grandit car le degré de liaison des variants génétiques avec le marqueur génotypé est moins certain. Parce que le déséquilibre de liaison (c'est-à-dire l'association préférentielle de génotypes) est moins conservé avec la distance, phaser des marqueurs éloignés est moins évident et nécessite d'avoir toutes les phases représentées dans la population de référence. La probabilité qu'une recombinaison ait lieu entre deux variants est fonction de la distance qui sépare ces variants. Plus grande est la distance plus grande est le risque que les allèles des deux marqueurs ne soient pas préférentiellement associés. L'algorithme d'imputation est donc plus à même d'estimer des phases erronées pour les individus à imputer lorsque la densité des typages est faible. Par exemple, Binsbergen *et al* (2014) obtiennent des qualités d'imputation bien meilleures lorsque l'imputation vers la séquence est faite depuis une puce HD qui compte plus de 777 000 marqueurs que depuis une puce de moyenne densité avec 50 000 marqueurs, et ce, quel que soit la taille de la population de référence utilisée (Van Binsbergen et al., 2014).

IV.4.c. Familles d'algorithmes d'imputation utilisées

Il existe plusieurs algorithmes d'imputation. Le choix d'un algorithme dépend des données et de la capacité de calcul disponibles. Les algorithmes peuvent être classés en deux catégories. L'imputation populationnelle s'appuie principalement sur le déséquilibre de liaison entre marqueurs proches pour construire une bibliothèque d'haplotypes à partir des génotypages (ce qui ne l'empêche pas de capturer indirectement de l'information familiale). La probabilité qu'un individu à imputer le soit avec un haplotype dépend des génotypes que ce dernier présente en basse densité et de la fréquence des haplotypes compatibles dans la population de référence. Après chaque imputation d'un individu, la bibliothèque et les fréquences des haplotypes sont mises à jour. Le principal inconvénient de cette méthode est qu'elle est généralement exigeante en termes de capacité et de temps de calcul. Elle est difficilement applicable sur des typages de très haute densité ou sur un grand nombre de typages. Elle permet toutefois d'imputer des populations sur lesquelles peu d'information d'apparentement est disponible (notamment en l'absence de pedigree). Des logiciels largement répandus utilisent cette méthode : Beagle (Browning & Browning, 2007), Minimac (Fuchsberger, Abecasis, & Hinds, 2015), ils permettent également d'obtenir un dosage allélique. Le dosage allélique est une variable continue comprise entre 0 (homozygote d'un allèle) et 2 (homozygote de l'autre allèle). Il permet de tenir compte de l'incertitude d'imputation dans les analyses qui suivent cette dernière.

Lorsqu'un pedigree est disponible l'imputation familiale est une alternative intéressante. L'utilisation du pedigree pour le phasage permet en effet d'imputer de façon plus fiable les génotypes des variants avec de petites MAF (Faux & Druet, 2017). Elle utilise les lois mendéliennes de transmission de l'information génétique et le pedigree d'un individu pour reconstituer son génotype à partir de celui de ses éventuels parents typés. L'imputation familiale intègre plus d'informations dans le choix des haplotypes et est, en cela, plus précise qu'une imputation basée uniquement sur les fréquences haplotypiques (Pausch et al., 2017). De plus, des parents aux descendants, les phases sont très peu modifiées car la recombinaison reste un évènement rare, on peut donc prédire un grand morceau des phases des descendants (long-range phasing) avec plus de certitude qu'en faisant appel à une imputation populationnelle. FImpute (Sargolzaei, Chesnais, & Schenkel, 2014) et AlphaImpute (Antolín, Nettelblad, Gorjanc, Money, & Hickey, 2017) sont des logiciels couramment utilisés qui font appel à cette méthode. AlphaImpute permet également de calculer des dosages alléliques.

V. La détection de QTL

Un QTL (Quantitative Trait Loci) est une région du génome qui est associée à un caractère quantitatif et qui explique une partie de la variabilité de ce caractère dans une population donnée. Leur recherche s'appuie classiquement sur des approches de type analyse de liaison, dans des dispositifs familiaux structurés, ou sur des analyses d'association quand la densité en marqueurs est suffisante. Dans la population caprine française de tels dispositifs ont été mis en œuvre et ont produit des résultats qui sont décrit ci-après.

V.1. Principe général de l'analyse de liaison

L'analyse de liaison s'appuie sur des dispositifs de génotypage familiaux. Dans les filières animales, ce sont souvent des dispositifs père-filles qui sont utilisés. L'analyse consiste à observer des co-ségrégations entre des allèles au sein même d'une famille et le cas échéant, une différence phénotypique entre les descendants portant l'une ou l'autre forme allélique. Dans le cas d'un dispositif père-fille, des morceaux complets de chromosomes sont transmis d'un père vers ses descendants directs. S'il y a eu des recombinaisons, elles sont peu fréquentes et éloignées (du fait de l'interférence notamment). Ainsi, un marqueur, même éloigné, s'il est transmis avec la mutation causale permettra d'identifier un QTL. Ce type d'analyse était donc largement utilisé lorsque la densité des marqueurs était faible (microsatellites notamment). Il est à noter que l'analyse de liaison suppose que la structure familiale de la population étudiée soit connue. De même, pour obtenir une puissance suffisante, l'analyse requiert un grand nombre d'individus génotypés intra-famille et un nombre suffisant de familles pour caractériser la diversité de la population. Bien que la méthode soit puissante, la localisation des QTL détectés est généralement peu précise car de grands segments de chromosomes sont transmis des parents au descendant.

V.2. Principe général de l'analyse d'association

A condition d'avoir suffisamment d'individus phénotypés et génotypés, l'analyse d'association pangénomique (ou GWAS pour Genome Wide Association Study) est une approche puissante pour détecter des régions qui contrôlent un phénotype. Elle est aujourd'hui la méthode prépondérante de détection de QTL, principalement parce que la densité des marqueurs sur les outils de génotypage est suffisante. Elle repose sur l'hypothèse d'un déséquilibre de liaison fort entre les marqueurs et les QTL.

L'approche s'appuie essentiellement sur des modèles mixtes polygéniques, tels que décrits ci-dessous :

$$\boldsymbol{Y} = 1\mu + \boldsymbol{x}b + \boldsymbol{u} + \boldsymbol{e}\ [1]$$

Où Y représente le phénotype du caractère; μ est la moyenne générale pour ce caractère dans la population; b est l'effet additif du marqueur ou variant testé; x est le vecteur des génotypes pour ce variant; u est le vecteur des effets polygéniques aléatoires, $u \sim N(0, G\sigma^2)$ avec G la matrice de parenté, calculée à partir des typages ou des pedigrees, elle permet d'établir la parenté des individus de l'analyse; e est le vecteur des effets aléatoires résiduels (effets non-captés par le reste du modèle) supposés normalement distribués. La matrice de parenté permet de prendre en compte la structure de la population et de corriger certain biais d'association entre un marqueur et le phénotype dû à des structures familiales particulières (si elles ne sont pas trop singulières).

Le modèle suppose que l'effet du génotype est additif, c'est-à-dire qu'il y a une relation linéaire entre l'effet du variant et le nombre de copies (0, 1 ou 2) d'un des allèles. On effectue ensuite un test de Student sur le paramètre b du modèle, pour tester la probabilité p que l'effet soit non-nul ($b \neq 0$). Le Manhattan plot est une représentation graphique de ces probabilités ($-\log(p)$) pour chacun des marqueurs du génome testés indépendamment (Figure 17). Un seuil de significativité à l'échelle du génome ou du chromosome est en général fixé pour tenir compte de la multiplicité des tests, par exemple par la correction de Bonferroni. Lorsque la p-value associée à l'effet d'un variant est inférieure à ce seuil, on considère que l'on a trouvé un variant d'intérêt.

Une approche haplotypique est également possible. Au lieu de tester l'effet du nombre de copies d'un allèle pour un variant donné, on évalue l'effet d'un haplotype. L'intérêt est de pouvoir capturer des variants rares présents dans certains haplotypes. La différence méthodologique est qu'on ne teste plus un facteur de régression b à l'aide d'un test T, mais un facteur a plusieurs niveaux d'effet non hiérarchisé (les haplotypes) par un test de rapport de vraisemblance. Parmi les différentes approches haplotypiques possibles, Druet & Georges (2010) ont proposé la constitutions d'états de phase. L'information locale de phase est codée sous forme numérique et peut ainsi être utilisée plus facilement dans les GWAS. Une valeur entière appelée état de phase est alors attribuée à chaque position du génotype. On teste ensuite son effet à l'aide du modèle suivant :

$$\boldsymbol{y} = W\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{e}$$

Dans ce modèle, y représente le vecteur des phénotypes pour le caractère étudié. β est le vecteur des effets des haplotypes. Les 2 haplotypes d'un individu sont deux effets aléatoires non-indépendants (l'un est niché dans l'autre). $\boldsymbol{u}$ est l'effet aléatoire de l'individu (effet génétique). W est une matrice d'incidence associée à l'effet haplotypique. $\boldsymbol{e}$ est l'effet aléatoire résiduel qui suit une loi normale telle que $e \sim N(0, I\sigma_e^2)$.

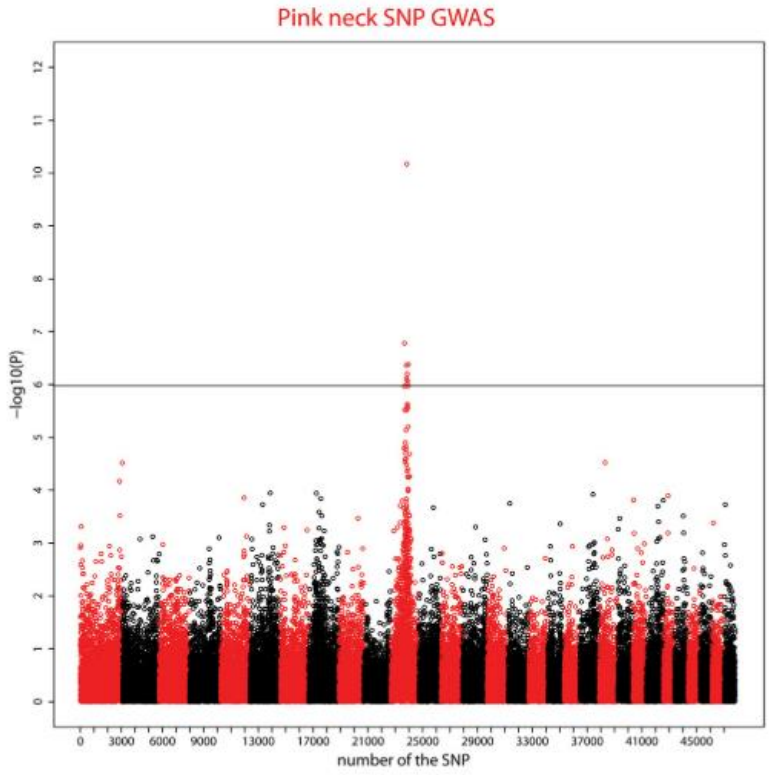


Figure 17: Exemple de Manhattan plot pour un phénotype de couleur indésirable de la robe en race Saanen française
(Martin et al., 2016)

V.3. Le dispositif QTL caprin

Les premiers génotypages 50k ont été acquis dans le cadre du projet européen 3SR (https://cordis.europa.eu/project/rcn/95054/factsheet/en) et le projet national PhénoFinLait qui a inclus les bovins, ovins et caprins laitiers avec pour objectif d'étudier la composition fine du lait (https://www6.inrae.fr/productions-animales/2014-Volume-27/Numero-4-PP-249-328/Avant-propos). L'objectif était de réaliser une première étude de détection de QTL sur les races Alpine et Saanen françaises. Cette première étude, initiée en 2011, s'appuyait sur un dispositif de familles de demi sœurs de père, dit petites-filles. Ainsi ont été retenues 20 familles issues de boucs d'insémination artificielle (11 Alpins et 9 Saanen) ayant eu au moins 100 filles nées entre 2008 et 2009. Ces boucs ont par ailleurs été séquencés et leurs séquences ont fait partie d'un des jeux de séquences qui ont permis la construction de la puce 50k

(Gwenola Tosser-Klopp et al., 2014). Leur choix répond à plusieurs critères. Leurs filles devaient être en première ou deuxième lactation pendant la phase de collectes des phénotypes du projet PhenoFinLait. Ces filles devaient être réparties parmi les 9 départements impliqués dans ce même projet. Les échantillons de sang et/ou des paillettes d'insémination des boucs en question devaient encore être disponibles dans le cas où des analyses supplémentaires devaient être faites. Enfin ces boucs devaient représenter au mieux la diversité génétique des deux races et être les moins apparentés possibles. Du côté des femelles, il a fallu sélectionner des élevages comportant des filles des 20 pères retenus. Finalement ce sont 209 élevages qui ont été retenus dans le cadre de projet comportant au total 4 500 filles issues des 20 boucs d'insémination. Finalement, ce sont 2 300 femelles des deux races qui ont été génotypées avec la puce caprine 50k dans le cadre de PhenoFinLait. Ces femelles possédaient des échantillons de sang dans le laboratoire de génotypage Labogena, et disposaient de phénotypes d'intérêt (morphologie mammaire, production laitière et comptages de cellules somatiques).

Aujourd'hui le grand nombre de boucs génotypés avec la puce 50k permet d'envisager des études d'association dans la population mâle, sans utiliser des structures de grande famille nécessaire à l'analyse de liaison. Les effectifs de reproducteurs mâles en testages sont bien moins nombreux qu'en bovins laitiers, mais les données s'accumulent. En effet, les mâles candidats au testage sur descendance sont systématiquement génotypés avec des effectifs en augmentation chaque année (Figure 18). Au total, en 2019, ce sont 4 025 génotypes 50k d'individus nés après 1993 qui ont nourri les travaux de cette thèse (Tableau 6).

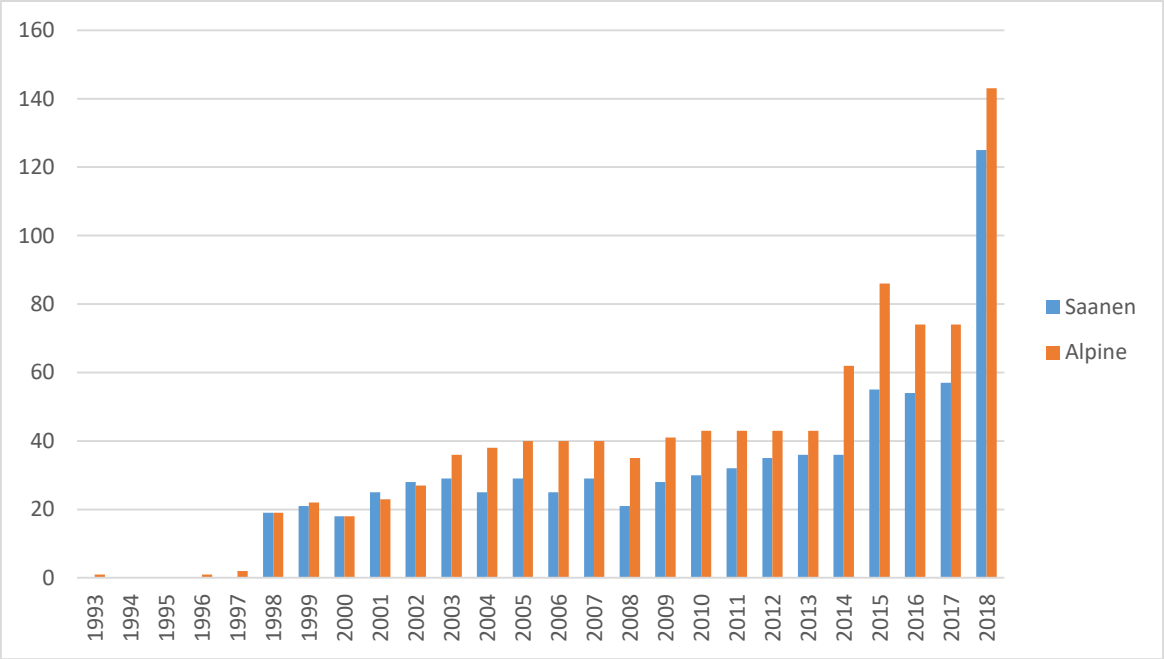


Figure 18: Répartition des 994 Alpins et 757 Saanen mâles génotypés par millésime de naissance

Tableau 5: Répartition des 4 025 individus génotypés avec la puce 50K en races Alpine et Saanen

	ALPINE		***SAANEN***	
	Mâles	Femelles	Mâles	Femelles
	994	1 461	757	813
TOTAL	*2 455*		1 570	

V.4. Les QTL précédemment identifiés en Alpine et Saanen avec la puce 50k

En Alpine et Saanen, le dispositif QTL a permis d'identifier deux mutations dans le gène DGAT1 (Diacylglycerol O-Acyltransferase 1) sur le chromosome 14 affectant le taux butyreux (Martin et al., 2017). Les deux mutations exoniques R251L et R396W sont, à ce jour, les seules à avoir été identifiées à l'aide d'analyses d'association conduites sur la puce 50k puis entièrement caractérisées par des analyses fonctionnelles. La première mutation est présente à hauteur de 3,5% dans la population Saanen alors que la seconde a été trouvée à des fréquences de 13 et 7% respectivement en Alpine et Saanen françaises. Les deux mutations

réduisent de façon significative le TB dans les 2 races en réduisant la synthèse des triglycérides par rapport à un génotype non-muté (sauvage).

La région du cluster des caséines sur le chromosome 6 (85,9 à 86,2 Mb sur la version ARS1 du génome caprin) était connue comme étant liée au taux protéique du lait de chèvre avant la création de la puce 50k (Martin & Leroux, 2000). Cette région a donc été densifiée en marqueurs afin de différencier au mieux les différentes combinaisons de mutations. Cependant, à l'heure actuelle, un typage spécifique à l'aide d'enzymes de restriction est encore nécessaire et effectué sur demande par le laboratoire Labogena pour différencier les génotypes pour la caséine alphaS1 (gène CSN1S1). La difficulté réside dans le fait que certains allèles sont déterminés par des épissages alternatifs ou de grandes insertions-délétions. Les 5 faux-sens qui permettent théoriquement de différencier les génotypes A, B1, B2, B3, B4, C et E (Figure 19) ont été sélectionnés pour la construction de sondes sur la puce. Toutefois, ils n'ont pas passés l'épreuve technologique de la construction des sondes ou ne permettent pas d'obtenir une qualité suffisante pour que le génotype CSN1S1 soit établi avec certitude.

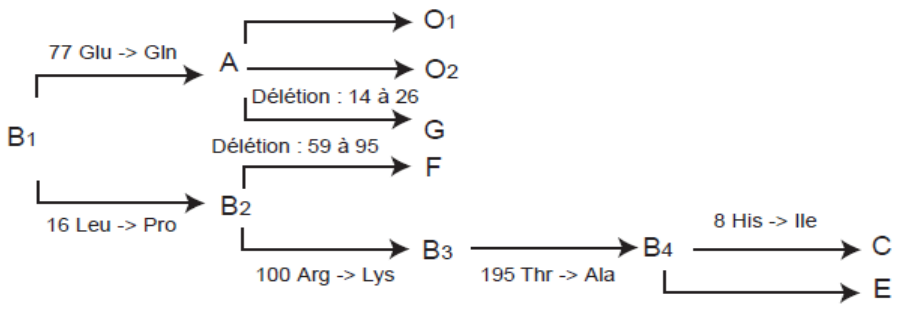


Figure 19: Phylogénie des différents génotypes au gène CSN1S1
(Martin & Leroux, 2000)

Des régions QTL ont été localisées pour d'autres caractères mais les mutations associées n'ont pour l'instant pas été finement caractérisées. En race Saanen, Martin *et al.* (Martin et al., 2018; Martin et al., 2017b) ont identifié une large zone pléiotropique sur le chromosome 19 associée à des caractères de morphologie et santé de la mamelle (LSCS) ainsi qu'à des caractères de production laitière. La région du signal est fortement enrichie en gènes avec une densité de 22,5 gènes par Mb. A titre de comparaison, d'après le NCBI (National Center for Biotechnology Information), la densité moyenne en gènes sur le génome caprin est

de 9,6 gènes par Mb. Plusieurs gènes candidats potentiels ont été désignés par Martin *et al*. pour la composition de lait car codant pour des protéines liées au métabolisme des acides gras: PLD2 (phospholipase D2), GGT6 (gamma-glutamyltransferase 6) et les protéines de la famille ALOX (arachidonate lipoxygenases) dont les gènes forment un cluster sur le chromosome 19. Toutefois, le lien de ces gènes avec les caractères de conformation de la mamelle n'est pas prouvé et n'est pas explicite. Des SNPs significatifs ont également été identifiés dans une race composite (Scottish Alpine, Saanen et Toggenburg) sur le chromosome 19 pour la production laitière et quelques caractères de conformation comme l'attache arrière, la profondeur de la mamelle, et les pattes antérieures (Mucha et al., 2018).

En Saanen, la présence de pampilles a également été étudiée en France (Martin, 2016) et en Suisse (Reber, Keller, Becker, Flury, & Welle, 2015). Des signaux ont été observé sur le chromosome 10 dans les deux études. Bien que des gènes candidats aient été proposés (FMN1 et/ou GREM1), la mutation causale n'a pas pu être identifiée car l'assemblage du génome de référence est encore incomplet dans la région (Reber et al., 2015).

En Alpine, des SNP liés à la qualité de l'attache arrière, le tour de poitrine et la forme de l'arrière-pis ont été localisés sur une région du chromosome 8 (Martin et al., 2018). Le même chromosome alpin présente un QTL associé au taux butyreux (Martin et al., 2017).

D'autres études ont été menées pour trouver des régions du génome associées à des tares. Les tares sont des défauts qui ne sont pas intégrés dans la sélection en tant que telle mais qui sont des causes d'élimination d'individus quelle que soit leurs valeurs génétiques sur les caractères en sélection. Leur définition est subjective et elles ne remettent pas toujours en question la viabilité de l'animal. En caprin, comme dans d'autres espèces laitières d'élevage, la présence de trayons surnuméraires qu'ils soient fonctionnels ou non, est une cause d'élimination systématique des mâles du schéma de sélection (environ 4,5% des boucs issus d'accouplement programmés en Saanen et 8,5% en Alpine). Une analyse d'association pour ce caractère a été réalisée sur les typages 50k, toutefois, aucune région précise n'a été identifiée (Martin et al., 2017). En Saanen, la couleur de robe rose est également considérée comme une tare car non-conforme au standard de la race (robe blanche uniforme). Sur les 15 dernières années, la présence de ce phénotype a été relevée dans la population française et elle concerne près de 20% des individus. Une analyse d'association a été conduite sur 810 femelles génotypées et a cette fois-ci révélé des zones significatives sur les chromosomes 5 et 13 (Figure 17). Le signal sur le chromosome 13 comporte le gène ASIP (Agouti Signaling Protein) qui est connu dans plusieurs espèces comme étant lié à la couleur de la peau ou du

pelage dont l'homme (Kanetsky et al., 2002), le cheval (Rieder, Taourit, Mariat, & Langlois, 2001) et le poisson rouge (Cerda, Haitina, Schio, & Peter, 2005).

VI. Les évaluations génomiques : principe et applications

Dans les années 70, le premier testage sur descendance et les premières évaluations génétiques ont été mis en place en France. Plus récemment, il y a environ 30 ans, les premiers outils permettant l'accès aux données moléculaires (marqueurs microsatellites) d'un individu sont entrés en scène. En caprins, le gène de la caséine alphaS1 a été identifié très tôt comme ayant un fort impact sur le taux protéique du lait (Grosclaude, Mahé, Brignon, Di Stasio, & Jeunet, 1987). Tous les mâles candidats au testage sur descendance sont typés spécifiquement pour ce gène, ce typage permet en particulier de choisir le meilleur candidat parmi des demi-frères. Le but est d'éliminer du socle de sélection tous les mâles porteurs d'allèles à l'effet délétère sur le taux protéique du lait (allèles O en particulier, E et F). De même, dans le cadre du plan d'élimination des allèles de susceptibilité à la tremblante, des typages sont effectués pour le gène spécifiant la protéine prion PRNP.

Depuis la mise à disposition de la puce 50k caprine, les mâles soumis au testage sur descendance sont génotypés dès leur plus jeune âge. La puce permet d'avoir une information mieux répartie sur le génome qu'un simple typage de gène (comme c'est le cas pour la caséine alphaS1). Le typage systématique des candidats au testage avec la puce 50k a permis de mettre en place les bases d'un dispositif permettant la recherche systématique d'association entre les phénotypes et les régions du génome dans la population actuelle. Dans les années à venir, tous les boucs candidats à la sélection seront typés avec la 50k qu'ils soient soumis par la suite à un testage sur descendance ou non (Figure 4).

La puce caprine a également permis l'instauration de la sélection génomique en caprins laitiers français en 2018. Cette dernière s'appuie sur la prédiction des valeurs génétiques des individus en sélection. Elle possède un intérêt économique : elle réduit le coût et le temps nécessaires au testage sur descendance en associant une valeur génétique à la naissance d'un individu. Toutefois, son principal enjeu est d'accélérer le rythme du progrès génétique. En effet, l'évaluation sur ascendants bien qu'elle soit elle aussi connue à la naissance d'un individu est moins précise que l'évaluation génomique. L'évaluation sur descendance reste nécessaire pour confirmer la valeur génétique d'un reproducteur, il faut cependant attendre qu'un mâle ait des filles en production pour l'obtenir. L'introduction de la génomique a modifié l'organisation du schéma de sélection. En effet, les candidats à la sélection sont désormais directement accouplés aux femelles du noyau de sélection sans attendre leur

confirmation par l'épreuve de la descendance (Figure 4). Ils participent donc dès leur plus jeune âge à l'amélioration de la race et à la création de la nouvelle génération de boucs améliorateurs.

Les évaluations génomiques des caprins laitiers français ont été étudiées au cours de différents travaux de thèse dont les conclusions sont résumées dans le paragraphe 6 de ce chapitre (Carillier-Jacquin et al., 2016; Carillier et al., 2013; Teissier et al., 2019; Teissier et al., 2018).

La faisabilité et la fiabilité des évaluations génomiques en races Alpine et Saanen ont été étudiées au cours des travaux de thèse de Céline Carillier (2012-2015), à partir des données de génotypes 50K disponibles et sur l'ensemble des caractères en sélection (caractères de production laitière, morphologie et scores de cellules somatiques). Ces travaux ont été conduits en utilisant les génotypes 50k de 470 Alpines et 353 Saanen alors disponibles.

Il a ainsi été établi que les conditions pour une évaluation génomique de qualité n'étaient pas entièrement réunies (Carillier et al., 2013) : la consanguinité et la persistance du déséquilibre de liaison (DL ou LD en anglais) sont faibles en caprins (consanguinité estimée sur génotypage d'environ 2% et LD de 0,17 entre 2 marqueurs consécutifs de la puce caprine), de plus peu d'animaux sont à la fois génotypés et phénotypés (environ 3000 pour les 2 races à l'heure actuelle). Ces derniers représentent toutefois correctement la structure et la diversité de la population du territoire français. Ainsi, il a été montré qu'une sélection génomique était faisable en caprins laitiers français sur la base d'un modèle multi-racial single-step GBLUP que nous détaillerons dans cette partie (Carillier, Larroque, & Robert-Granié, 2017).

VI.1. Principe de l'évaluation génomique

Les évaluations génomiques s'appuient sur la construction d'une population de référence pour une race ou une espèce donnée (dans ce cas, on parlera d'évaluation multiraciale). Cette population de référence est constituée d'animaux génotypée et phénotypée. Elle permet l'établissement d'un lien statistique entre le génotype d'un individu aux marqueurs génotypés et ses performances ou les performances de ses descendants. Ce lien statistique permet par la suite de prédire une valeur génétique pour des individus n'ayant pas encore de performance propre ou de descendants avec performance mais étant génotypés. Il est estimé que pour obtenir une relation fiable entre génotype et phénotype, il faut plus de 1 000 individus dans la population de référence. La taille de cette population est également à

adapter à l'héritabilité du caractère considéré. Moins le caractère est héritable et plus la population de référence doit être de taille importante. (Robert-Granié, Legarra, & Ducrocq, 2011)

Pour estimer la qualité des prédictions, on divise une population de mâles pour lesquels on dispose de filles avec des performances enregistrées. Une partie de ces derniers, les plus anciens, constituera la population de référence qui servira à calibrer le modèle. Les animaux les plus récent constitueront la population de validation pour laquelle on estimera une valeur génétique (EBV) sans utiliser les performances de leurs filles. Ces valeurs génétiques sont ensuite comparées aux performances réalisées (DYD). On estime ainsi une précision des évaluations en calculant une corrélation entre les EBV et les DYD et un biais à l'aide d'une régression linéaire entre EBV et DYD.

VI.2. Les modèles « généraux » d'évaluations génétique et génomique

Les évaluations génétiques comme génomiques reposent sur le postulat que le phénotype observé est la résultante d'un effet génétique et d'effets d'environnement. Les modèles linéaires mixtes qui en découlent suivent donc ce principe :

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{e} \text{ [2]}$$

Où y représente le vecteur des observations ou phénotypes. X et Z sont les matrices d'incidence qui relient les différents vecteurs du modèle. u est le vecteur aléatoire des effets génétiques, il suit une loi normale centrée sur 0 et de variance $A\sigma_u^2$. A représente la matrice de parenté , estimé à partir de la généalogie (ou pedigree) des individus considérés. e est le vecteur des résidus distribué selon une loi normale d'espérance nulle et de variance $I\sigma_e^2$.

β est le vecteur des effets fixes ou effets d'environnement dont on sait qu'ils influencent significativement le phénotype étudié. Ces effets peuvent être fixes ou aléatoires. En général, lorsqu'un effet a un nombre limité de valeurs différentes, on considère que c'est un effet fixe. Dans le cas contraire, on préfère imposer que l'effet est aléatoire, l'estimation des effets des différents niveaux n'étant pas nécessaire. L'identification exhaustive des paramètres influençant le caractère étudié est capitale, en cas d'oubli d'un facteur, les résultats des évaluations peuvent être biaisés.

Classiquement, pour résoudre ce modèle et estimer les valeurs des différents effets, le BLUP (Best Linear Unbiaised Prediction) modèle animal est utilisé. Henderson *et al.*

(Henderson, Kempthorne, Searle, & Krosigk, 1959) ont dérivé un système d'équations qui en permettent la résolution :

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Dans ce système A^{-1} est l'inverse de la matrice de parenté, σ_u^2 la variance génétique et σ_e^2 la variance résiduelle. En résolvant ce système, il est possible d'estimer simultanément et de façon non-biaisée les effets fixes et aléatoires. (Henderson, 1975)

L'évaluation génomique cherche à attribuer une valeur à un candidat à la sélection dès son plus jeune âge. On ne dispose alors pas encore de performance propre ou de performance pour ses descendants mais seulement d'un génotype pour cet individu. Cette valeur est appelée GEBV (Genomic Estimated Breeding Value) et correspond à la capacité qu'a cet individu à transmettre un potentiel génétique à ses descendants. Un des premiers modèles développés, appelé BLUP-SNP, pour calculer les GEBV est le suivant :

$$\boldsymbol{y} = \mathbf{1}\mu + X\boldsymbol{g} + \boldsymbol{e} \text{ [3]}$$

Dans lequel y est le vecteur des phénotypes préalablement corrigés pour tous les effets de milieu identifiés précédemment ; g est le vecteur des effets de chacun des marqueurs considérés ; X est la matrice des génotypes de chaque individu pour chacun des marqueurs et enfin e est le vecteur des résiduelles. La valeur génétique d'un animal est alors estimée à partir des n marqueurs du génotypage :

$$GEBV_i = \sum_{j=1}^{n} x_{ij} g_j \text{ [4]}$$

Où g_j est l'effet du SNP j ; x_{ij} est le génotype de l'individu i pour le SNP j.

Dans la majorité des cas, on suppose que chacun des SNP a un effet faible sur le phénotype étudié. De plus, on considère que l'ensemble des effets suit une loi normale centrée sur 0 et de variance σ_g^2. Enfin, les effets sont supposés indépendants.

Ce modèle [3] est robuste. Un autre modèle génomique, appelé GBLUP, se base sur le même modèle défini en [2] en remplaçant la matrice de parenté **A** par une matrice **G**, construite à partir des informations des SNPS. En caprins, le passage d'un modèle d'évaluations basé sur le pedigree (BLUP) à un modèle génomique (GBLUP) a permis d'obtenir des gains compris entre 3,4% (TP) et 21,3% (avpis). Néanmoins, il connait quelques limites notamment en ce qui concerne les gènes majeurs pour lesquels l'effet de quelques SNP est très fort (Robert-Granié et al., 2011). Plusieurs variantes ont alors été proposés dans la littérature.

VI.3. Quelques variantes aux modèles traditionnels

Il existe de nombreuses alternatives aux modèlex BLUP-SNP et GBLUP. Le *gene content* permet ainsi d'intégrer le génotype pour une mutation à très fort effet dans un modèle multi-caractère (Gengler, Mayeres, & Szydlowski, 2007; Legarra & Vitezica, 2015). Enfin, plusieurs méthodes bayésiennes ont été développées. Les différentes approches bayésiennes, proposées dans la littérature, se distinguent les unes des autres par les hypothèses qu'elles impliquent sur la distribution des effets des marqueurs (B. J. Hayes et al., 2014).

Pour pallier le fait que le BLUP-SNP nécessite un grand nombre de typages pour que l'on puisse estimer correctement les effets des marqueurs, un modèle single-step GBLUP (ou ssGBLUP) a été développé (Aguilar et al., 2010). Ce dernier permet en une seule étape de tenir compte à la fois des génotypages pour les individus qui en possèdent mais aussi des informations du pedigree quand elles sont disponibles. Dans le cas du single-step GBLUP (ssGBLUP), la matrice de parenté (A) des équations précédentes (modèle 2) est remplacée par une matrice H qui combine l'information de parenté généalogique (A) et l'information de parenté génomique fournie par le génotypage d'un sous-ensemble de la population. La matrice H se définit assez simplement de la façon suivante :

$$H^{-1} = A^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{pmatrix} \quad [5]$$

A_{22}^{-1} est ici l'inverse de la matrice de parenté généalogique pour les animaux génotypés. (Aguilar et al., 2010).

Une des assertions du modèle général GBLUP et du modèle ssGBLUP consiste à supposer que tous les variants inclus dans le modèle ont un effet faible et que c'est la combinaison de tous ces effets faibles qui explique les variations à l'échelle de l'individu. Cette hypothèse est contraignante et conduit à des erreurs car en réalité tous les marqueurs

n'ont pas un effet équivalent et parfois certains ont des effets majeurs (cas des gènes majeurs notamment). Pour remédier à cela plusieurs approches ont été développées pour répondre à des cas particuliers.

VI.3.a. Le Weighted ssGBLUP classique

Une des stratégies alternatives proposées consiste à accorder des poids aux différents variants inclus dans le modèle en fonction de la magnitude de leur effet sur le caractère étudié. Cette méthode est communément appelée Weighted single-step GBLUP (WssGBLUP) et est un processus itératif. Un premier modèle comparable à un ssGBLUP classique est mis en oeuvre sur les données pour estimer les effets des variants. Ces derniers sont ensuite convertis en poids qui sont utilisés et ré-estimés à chacune des itérations suivantes. L'intégration des poids se fait dans le calcul de la matrice G^* à l'aide de la formule suivante (VanRaden, 2008) :

$$G^* = 0{,}95 * \frac{ZDZ'}{\sum_{i=1}^{m} p_i(1-p_i)} + 0{,}05 * A_{22} \quad [6]$$

Où Z est la matrice des génotypes corrigés par la fréquence des allèles des SNPs. D est la matrice des poids des différents variants (à la première itération D est la matrice identité). m est le nombre total de variants. p_i est la fréquence des allèles du variant. A_{22} est la matrice de parenté basée sur le pedigree.

Les étapes ont été décrites par (Wang et al., 2014; Zhang et al., 2016) comme les suivantes :
Où $\lambda = \frac{1}{\sum_{i=1}^{m} p_i(1-p_i)}$.

1 Première itération : phase d'initialisation
$D_{(1)} = I$; $G^* = 0{,}95 * \lambda Z D_{(1)} Z' + 0{,}05 * A_{22}$
2 Calcul de G^* avec la formule ci-dessus pour obtenir le vecteur des valeurs génétiques $\hat{u}_g$
3 Passage à l'itération 2
4 Estimation des effets des variants $\hat{a}_{(it)} = \lambda D_{(it-1)} Z' G^{*-1}_{(it-1)} \hat{u}_{g\,(it-1)}$
5 Conversion des effets pour obtenir les poids des variants. Pour chaque variant on calcule donc un poids selon la formule : $d_i^* = \hat{a}_i^2 * 2p_i(1-p_i)$
Ces poids sont ensuite intégrés à la matrice $D^*_{(it)}$.

6 Normalisation des poids $D_{(it)} = \frac{tr(D_{(1)})}{tr(D^*_{(it)})} * D^*_{(it)}$

7 Construction de la matrice $G^*_{(it)} = 0{,}95 * \lambda Z D_{(it)} Z' + 0{,}05 * A_{22}$

8 Lancement d'un WssGBLUP intégrant $G^*_{(it)}$ pour obtenir le nouveau vecteur des valeurs génétiques $\hat{u}_{g\ (it)}$

9 Passage à l'itération suivante $it = it + 1$

10 Arrêt du processus itératif ou nouveau lancement à partir de l'étape 4

Ce modèle a toutefois tendance à s'emballer en écrasant l'effet des variants qui avaient un poids initial faible et en augmentant les poids des variants dont l'effet estimé initialement était fort. Ce faisant, au fur et mesure des itérations, la qualité des évaluations est dégradée. Zhang et al. (2016) signalent d'ailleurs qu'il est préférable de s'arrêter à la 2ème itération pour limiter les défauts.

VI.3.b. Le Weighted ssGBLUP avec des fenêtres de variants consécutifs

Pour limiter aux défauts de la méthode précédente, Zhang et al. (2016) propose, à partir de la première itération d'un WssGBLUP, d'attribuer le même poids à plusieurs SNP consécutifs en construisant des fenêtres non-chevauchantes de plusieurs variants. Le poids est alors attribué à une fenêtre et non plus à un seul SNP ce qui évite que l'effet de certains SNP soit complétement écrasé par les effets plus forts d'autres marqueurs. Ce poids est la somme, la moyenne ou le maximum des poids des variants de la fenêtre. Ces approches se sont avérées relativement efficaces pour la construction de Manhattan plots et l'identification de régions d'intérêt car le bruit était réduit (Wang et al., 2014). Ceci a été particulièrement vrai sur données simulées lorsque le caractère était influencé par peu de régions différentes du génome (Zhang et al., 2016).

Marc Teissier s'est attaché pendant sa thèse (2015-2018) à appliquer des modèles de WssGBLUP aux données caprines de race Alpine et Saanen (Marc Teissier et al., 2018). Les gains ont varié entre 5 et 14% en fonction de la race ou du caractère considéré. En Saanen où des QTLs ont été détectés sur plusieurs caractère, cette approche s'est avérée particulièrement intéressante. En Alpine, c'est principalement l'évaluation du taux protéique qui a été améliorée car le caractère est fortement associé à la région des caséines sur le chromosome 6. Pour les caractères polygéniques, un ssGBLUP est suffisant pour prédire correctement les valeurs génétiques des individus.

VI.4. L'intégration de données de séquence aux évaluations génomiques

Les données de séquence à grande échelle sont intéressantes en cela qu'elles contiennent en théorie les mutations causales qui expliquent les variations d'un caractère donné dans la population étudiée. Les évaluations génomiques actuelles se font en général sur des puces de moyenne densité qui, la plupart du temps, n'incluent pas directement les mutations. Ces évaluations reposent donc sur le déséquilibre de liaison entre les marqueurs de la puce et les mutations causales (Figure 20).

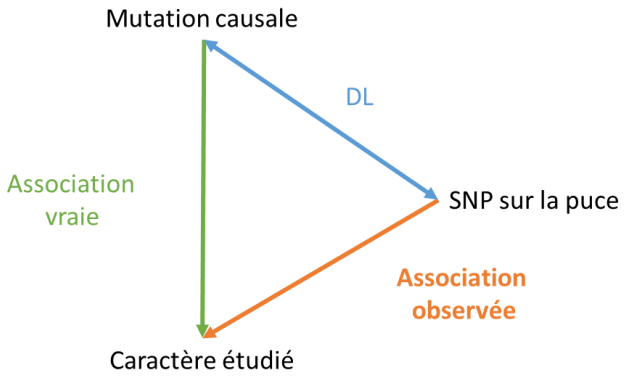


Figure 20: Principe de l'utilisation de données de génotypages pour l'évaluation génomique
adapté de (Legarra, 2014)

Plusieurs méthodes ont été testées pour intégrer les données de séquence aux évaluations génomiques. Les données de séquence se caractérisent par un très grand nombre de variants (en général plusieurs millions voire dizaines de millions). Ainsi, très peu d'études se sont attachées à intégrer l'ensemble des variants tel quel dans les modèles car les temps de calcul s'en trouvent rallongés ce qui est un handicap pour une application en routine. Toutefois, en ovins (Moghaddar et al., 2018) et en bovins laitiers (B. J. Hayes et al., 2014), des études ont démontré des gains de précision en intégrant les données de séquence au calcul de la matrice de parenté génomique compris entre 1,4% et 2,6% par rapport à des typages 50k en ovins et 2% par rapport à des typages HD (800k) en bovins laitiers.

Les données de séquence comprennent plusieurs millions de variants, il est donc préférable pour limiter les temps de calculs de sélectionner un sous-ensemble des variants de la séquence avant de les ajouter aux SNP disponibles sur les typages d'une puce pour les évaluations génomiques. Les critères de sélection sont multiples. Ainsi, il est possible de retenir les variants répondant à un ou plusieurs des critères suivants (Brøndum et al., 2015; Hayes et al., 2014; Moghaddar et al., 2018; VanRaden et al., 2017) :

- Significatifs sur les analyses d'association. Il faut dans l'idéal que ces analyses aient été conduites sur un sous-ensemble d'animaux indépendants de ceux qui servent à l'évaluation pour ne pas introduire un biais. Parfois, un filtre sur le déséquilibre de liaison entre les variants significatifs est appliqué pour n'introduire que des variants indépendants les uns des autres dans les évaluations.
- Fort effet estimé dans les évaluations génomiques. Dans ce cas de premières évaluations à partir de l'ensemble des marqueurs sont nécessaires.
- Forte part de variance expliquée. Cette part est calculée comme $2p(1-p)\alpha^2$ avec p la fréquence d'un des allèles du marqueur et α son effet estimé
- Candidats potentiels (après vérification de l'annotation pour ce marqueur). L'annotation est alors extraite et on choisit les variants en fonction du type de mutation (faux-sens, synonyme…) ou de sa position par rapport à un gène d'intérêt.
- Nature du variant (SNP ou indel). Les indels sont supposés avoir un effet bien plus fort car ils décalent complétement le cadre de lecture pour la traduction, toutefois, leur alignement sur le génome et donc leur qualité est à vérifier précautionneusement.
- Distance à un marqueur de la puce utilisée pour les évaluations. L'objectif en ajoutant des marqueurs dans le modèle d'évaluation génomique est d'améliorer la précision des valeurs génétiques qui en sortent. L'idéal est de couvrir le génome en complétant les génotypages issus des puces, il est donc préférable de sélectionner des variants de séquence qui ne soient pas trop proches des marqueurs de la puce.

Certains auteurs ont trouvé intéressant de décomposer le terme $X\boldsymbol{g}$ du modèle général [3] pour différencier les marqueurs de la puce de ceux de la séquence. On obtient ainsi le modèle suivant :

$$\boldsymbol{y} = \mathbf{1}\mu + X_{puce}\boldsymbol{g_{puce}} + X_{seq}\boldsymbol{g_{seq}} + \boldsymbol{e} \text{ [7]}$$

Les résultats de cette décomposition sont variables. Brøndum et al. (2015) ont, en effet, obtenu une meilleure précision des estimations lorsque les marqueurs associés à des QTL étaient traités séparément de ceux de la puce 54k pour la majorité des caractères qu'ils ont étudiés en bovins laitiers (R F Brøndum et al., 2015). En revanche, en ovins, bien que l'intégration d'un deuxième terme pour les marqueurs de séquence améliore la précision des évaluations par rapport à la puce 50k (+4,4% ou +3,8% en fonction de la race), Moghaddar et

al. (2018) parviennent à un gain plus élevé avec une seule variance pour les 2 types de marqueurs (+6,2% ou +4,1% pour les mêmes races).

VII. Objectifs de la thèse

Dans ce chapitre, nous avons vu comment s'organise la filière caprine laitière en France, comment est effectuée la sélection et quels sont les caractères d'intérêt. Nous avons également souligné l'intérêt du séquençage pour obtenir des données génétiques exhaustives pour un individu. La filière caprine française a intégré l'ère de la génomique avec le récent développement et la valorisation d'une puce à ADN dans les années 2010-2020 pour la recherche de QTL et l'évaluation génétique. La démocratisation des données de séquençage tout génome pour les animaux de rente ouvre de nouvelles perspectives. Le projet VarGoats nous a fourni 40 séquences en race Alpine et 33 en race Saanen. Le nombre d'individus dans les deux races françaises est faible. L'imputation directe depuis la seule puce disponible en caprins représente donc un challenge, d'autant plus que la diversité génétique dans les races est grande. Les méthodes et résultats publiés en bovins ne seront donc pas directement transposables à nos races d'étude. L'étude de la qualité d'imputation vers la séquence dans la filière caprine est un préalable nécessaire à l'utilisation de cette dernière dans le dispositif de détection de QTL ainsi que dans les évaluations génomiques.

L'objectif principal de ma thèse est d'étudier l'intérêt de l'intégration des données de séquence dans la filière laitière caprine française. La mise en place d'un contrôle de la qualité des données de séquence a représenté un travail majeur dans ma thèse. Il s'est appuyé sur une recherche bibliographique ainsi que sur la comparaison des génotypes 50k disponibles avec les séquences filtrées.

Une étude préalable de l'imputation depuis la puce 50k vers la séquence a ensuite été mise en œuvre dans le but d'obtenir un nombre suffisant de séquences imputées de bonne qualité. Plusieurs méthodes d'imputation (imputation populationnelle ou familiale) et plusieurs logiciels ont été testés en utilisant les données de séquence disponibles (829 séquences).

Les séquences imputées des mâles ont permis la confirmation de QTL précédemment observés sur les génotypes 50k ainsi que la détection de nouvelles régions d'intérêt. La densité des données de séquence représentait une opportunité sans précédent d'approfondir une région QTL du chromosome 19 en Saanen qui est associée à la fois à des caractères de

production mais aussi à des caractères de morphologie et santé de la mamelle ainsi qu'à des caractères de production de semence.

Enfin, en réunissant l'ensemble des travaux effectués précédemment nous avons étudié l'impact de l'intégration de données de séquence imputées sur le chromosome 19 sur la précision des évaluations en race Saanen françaises. Plusieurs modèles d'évaluations ont pu être alors comparés : ssGBLUP, WssGBLUP en utilisant différents panels de variants imputés. Cette analyse nous a permis d'ouvrir des pistes de réflexion pour améliorer la précision des évaluations en race Saanen en valorisant l'information provenant de séquences complètes et de proposer des perspectives d'études complémentaires.

Chapitre 2
Contrôle qualité des données et imputation vers la séquence

Dans cette partie, nous ferons une description détaillée des données de séquence mises à notre disposition par le consortium VarGoats. Un sous-jeu de 829 séquences a été utilisé pour en implémenter un filtrage solide. Ce filtrage a été éprouvé par analyse des résultats d'imputation et d'analyses d'association qui l'ont suivi. Une fois le filtrage fixé, nous avons pu définir la meilleure stratégie d'imputation en caprins laitiers français et mesurer son impact sur la détection de QTL.

Nous présenterons dans un premier temps les données de séquences acquises dans le cadre du consortium VarGoats. Ces données ont ensuite fait l'objet d'un contrôle qualité. Enfin, nous avons défini la stratégie optimale pour maximiser la qualité d'imputation. La capacité des données imputées à détecter des régions d'intérêt sur le génome de QTL a fait l'objet d'une attention particulière. Cette partie comporte un *data paper* soumis le 10/06/2020 à *GigaScience* ainsi qu'un article de recherche publié dans *BMC Genetics* le 21/02/2020. Des travaux additionnels non-publiés sont également présentés pour développer le cheminement qui nous a amené à définir le filtrage et la méthode d'imputation choisie.

I. Les données de séquence du projet VarGoats – Article

I.1. Introduction et résumé de l'article

Depuis l'arrivée sur le marché des séquenceurs haut-débit dans la deuxième moitié des années 2000, les données de séquences sont devenues plus accessibles. En réduisant le temps d'acquisition des séquences, les séquenceurs haut-débit ont permis de grandes économies d'échelle. La première espèce a en avoir bénéficié est l'espèce humaine pour laquelle un premier projet 1000 génomes (ou 1000G) a débuté en 2008 (Auton et al., 2015). Par la suite dans les espèce d'élevage d'autres projets de grande envergure sont apparus : 1 000 génomes bovins en 2012, 1 000 génomes ovins en 2016, etc… (Bolormaa et al., 2019; Ben J Hayes & Daetwyler, 2019)

En caprins, en 2010, la première séquence profonde d'une femelle de race Black Yunnan a été assemblée par une équipe de chercheurs en Chine (assemblage CHIR1.0) (W. Wang et al., 2013). Cette séquence a initialement servi de référence pour l'espèce avant l'arrivée d'un nouvel assemblage d'une séquence d'un mâle de race San Clemente aux Etats-Unis (assemblage ARS1) (Bickhart et al., 2017). Elle a abouti *in fine* à la création de la première puce à ADN caprine, disponible en 2011 (Gwenola Tosser-Klopp et al., 2014). Un total de 52 295 loci a été sélectionné sur le génome caprin dans 5 gene pools : Alpine, Boer, Creole, Katjang/Savanna and Saanen.

Les pères du dispositif de détection de QTL en Alpine et Saanen françaises ont été séquencés et ont notamment contribué à la sélection des SNPs pour la création de la puce caprine. Suite à cette première vague de séquençage, un projet 1 000 génomes a été lancé par l'IGGC (International Goat Genome Consortium), le projet VarGoats (http://www.goatgenome.org/vargoats.html). Sur la base des super-populations et populations définies dans le cadre du projet AdaptMap (Evol et al., 2018), des individus ont été sélectionnés pour représenter leur race. Le *data paper* ci-après décrit précisément la nature et la composition des données de séquence disponibles dans l'espèce caprine.

Le jeu de séquences caprines définitif comprend 1 160 séquences. Parmi ces dernières, on compte 652 nouvelles séquences, 291 séquences publiques et 217 séquences issues du projet NextGen. Ces données incluent 35 animaux appartenant à 8 espèces sauvages. Cet ensemble d'individus a permis d'identifier 74 274 427 SNPs et 13 607 850 suite à un alignement sur la dernière version du génome caprin (ARS1). Des analyses phylogénétiques montrent que les animaux africains, européens et asiatiques forment des clusters indépendants. Les caprins d'Océanie et des Caraïbes (race Créole) se répartissent sur l'arbre en fonction des animaux d'importation dont ils sont issus.

Ce jeu de données sera la base d'études des processus de sélection liées à la domestication. Il sera également une ressource précieuse quant à la recherche de polymorphismes causaux de caractères complexes. L'ensemble de ces travaux sera assuré par des groupes du consortium VarGoats (http://www.goatgenome.org/vargoats_workgroups.html).

I.2. Le projet VarGoats, un jeu de 1 160 séquences complètes pour analyser la diversité mondiale de l'espèce *Capra hircus* : Article

VarGoats project: a 1,160 whole-genome sequence dataset to dissect *Capra hircus* global diversity

Laure Denoyelle[1,2,†] (https://orcid.org/0000-0001-6343-7398), Estelle Talouarn[1,†] (https://orcid.org/0000-0002-5016-0446), Philippe Bardou[1,3] (https://orcid.org/0000-0002-0004-0251), Licia Colli[4] (https://orcid.org/0000-0002-7221-2905), Adriana Alberti[5] (https://orcid.org/0000-0003-3372-9423), Coralie Danchin[6] (https://orcid.org/0000-0002-0671-2792), Marcello Del Corvo[4], Stéfan Engelen[5] (https://orcid.org/0000-0003-0003-1192), Céline Orvain[5], Isabelle Palhière[1], Rachel Rupp[1] (https://orcid.org/0000-0003-3375-5816), Julien Sarry[1], Mazdak Salavati[7,8] (https://orcid.org/0000-0002-7349-2451), Marcel Amills[9] (https://orcid.org/0000-0002-8999-0770), Emily Clark[7,8] (https://orcid.org/0000-0002-9550-7407), Paola Crepaldi[10] (https://orcid.org/0000-0002-6526-2162), Thomas Faraut[1] (https://orcid.org/0000-0001-5156-3434), Clet Wandui Masiga[11], François Pompanon[2] (https://orcid.org/0000-0003-4600-0172), Benjamin D Rosen[12](https://orcid.org/0000-0001-9395-8346), Alessandra Stella[13] (https://orcid.org/0000-0003-2850-3964), Curtis P Van Tassell[12] (https://orcid.org/0000-0002-8416-2087), Gwenola Tosser-Klopp[1,*] (https://orcid.org/0000-0003-0550-4673) and the VarGoats Consortium[#]

Institutions:

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

[2] Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LECA, 38000 Grenoble, France

[3] Sigenae, INRAE, 31326 Castanet Tolosan, France

[4] Dipartimento di Scienze Animali, della Nutrizione e degli Alimenti, BioDNA Centro di Ricerca sulla Biodiversità e sul DNA Antico, Università Cattolica del Sacro Cuore, Piacenza, 29122 Italy

[5] Génoscope, Institut François Jacob, CEA, CNRS, Université Paris-Saclay, Université Evry Val-d'Essonne, Evry, France

[6] Institut de l'Elevage, Maison Nationale des Eleveurs - 149 Rue de Bercy - 75595 Paris cedex 12, France

[7] The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush Campus, EH25 9RG, UK

[8] Centre for Tropical Livestock Genetics and Health (CTLGH), Easter Bush Campus, EH25 9RG, UK

[9] Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

[10] Depth. Agricultural and Environmental Sciences - Production, Landscape, Agroenergy, University of Milan, Milan, Italy

[11] Tropical Institute of Development Innovations (TRIDI), P O Box 23158, Kampala, Uganda

[12] Animal Genomics and Improvement Laboratory, BARC, USDA-ARS, Beltsville, MD 20705, USA

[13] Istituto di Biologia e Biotecnologia Agraria, Consiglio Nazionale delle Ricerche, Milan, Italy

Correspondence:

[*] Corresponding author: gwenola.tosser@inrae.fr

Authors' Contribution:

[†] LD and ET contributed equally to the work

[#] The list of sample providers and their affiliations is given at the end of this paper and a full list of working group participants is available here: http://www.goatgenome.org/vargoats_workgroups.html

Abstract

Background

Since their domestication 10,500 years ago, goat populations with distinctive genetic backgrounds have adapted to a broad variety of environments and breeding conditions. The VarGoats project is an international 1,000 genomes resequencing program designed to understand the consequences of domestication and breeding on the genetic diversity of domestic goats, as well as to elucidate how speciation and hybridization have modeled the genomes of a set of species representative of the genus *Capra*.

Findings

A dataset comprising 652 sequenced goats and 508 public goat sequences, including 35 animals from 8 wild species, has been collected worldwide. We identified 74,274,427 SNPs and 13,607,850 INDELs when aligning sequences to the latest version of the goat reference genome (ARS1). By performing a phylogenetic analysis, we have found that goats from Africa, Asia and Europe tend to group in independent clusters. Since goat breeds from Oceania and Caribbean (Creole) all derive from importations, they are distributed along the tree according to their ancestral geographic provenance.

Conclusions

Here we report an unprecedented international effort to characterize the genome-wide diversity of domestic goats. This wide panel of sequenced individuals represents an opportunity to ascertain how the demographic and selection processes associated with post-domestication history have shaped the diversity of this species. Data generated in the project might be also extraordinarily useful to identify deleterious mutations as well as polymorphisms with causal effects on complex traits, thus generating new knowledge that could be used in genomic prediction and genome-wide association studies.

Keywords

Whole-genome sequencing, Goat, *Capra*, Genomic variation, Genetic diversity

Data description

Context

Goats (*Capra hircus*) were domesticated around 10,500 years ago in the Fertile Crescent [1]. Since then, goat populations have gradually specialized and now provide milk, meat, fibers [2] and fuel from manure, to become an important pillar of the livestock sector in many countries around the world. According to FAOSTAT, (www.fao.org/faostat/en) the world goat population increased by 38% since 1994, reaching over 1.03 billion heads in 2017. This makes goats the 3rd most important ruminant production species. Goats have undergone an intense process of adaptation and occupy several diverse agroecological regions of the world.

In 2010, the first goat reference whole-genome sequence was assembled [3] and the International Goat Genome Consortium (IGGC) was created to further support the development of genomic tools for studying the genetic variation of domestic goats. Later on, a 50k-SNP (Single Nucleotide Polymorphism) panel, the GoatSNP50 BeadChip was developed [4] to facilitate QTL discovery through genome-wide association studies [5–7] and to enable genomic selection [8,9]. This resource became a tool that facilitated collaboration, because data from diverse sources were interchangeable. The ADAPTmap project [10–16], was one such collaboration, compiling goat genotypes from across the globe and exploring genetic diversity. However, the ADAPTmap dataset was limited to a subset of countries, and did not fully represent the variability of the *Capra* species worldwide. Indeed, wild goat species other than the bezoar (*Capra aegagrus*) were not investigated. Moreover, data generated from SNP chips are known to be affected by ascertainment bias [17], a limitation that can be overcome by carefully filtering whole genome sequencing data of sufficient depth [18]. Here we report an

international resequencing effort: the VarGoats project, which has generated a data set of 1,160 goat genomes through the generation of new sequencing data as well as by retrieving existing genome sequences from public databases. The VarGoats dataset will pave the way towards obtaining an unprecedented perspective about the natural and human-mediated evolutionary forces which have shaped genomes and the genetic diversity of domestic goats.

Ethics Statement

Blood collection or ear-tags samples were carried out in accordance with the national regulations from the countries where such samples were collected. In the case of samples sequenced at Génoscope (Evry, France), DNA was imported into France either with authorization 31 555 50, delivered on May, the 24th 2016 for European countries or covered by an import permit from DDPP (Direction Départementale de la Protection des Populations) for non-European countries. The only exception was the African animals for which sequencing was performed at Edinburgh Genomics. The DNA for these libraries was imported into Scotland by permission of the Scottish Government Animal Health and Welfare Division and the UK under generic license IMP-GEN-2008-03.

Individual selection

Animals sequenced in the VarGoats project were selected to represent the international genetic diversity in goats. We selected 468 animals already included in ADAPTmap dataset (with 457 public genotypes) and added 184 animals from different breeds and locations in order to improve the representation of the current levels of genetic diversity in goats. After sequencing all samples (652), 217 additional samples were retrieved from NextGen Consortium projects (PRJEB3134, PRJEB3135, PRJEB4371, PRJEB5166, PRJEB3136 and PRJEB5900 studies) and 291 additional sequences from public

sequence data repositories (extraction from NCBI database on the 2019/02/07) were added to the overall VarGoats data set.

However, some individuals were selected for specific research purposes. This is the reason why we can observe a distortion in the number of sequenced Alpine and Saanen for instance. Besides some individuals can be closely related, an analysis is therefore advised to anyone using the data for any analysis.

Among the 1,160 animals, 35 represented wild *Capra* species (3% of the dataset, *C. aegagrus*, *Capra caucasica*, *Capra cylindricornis*, *Capra falconeri*, *Capra ibex*, *Capra nubiana*, *Capra pyrenaica* and *Capra sibirica*) distributed with domestic goats native to Africa (450 goats, 39%), Europe (443 goats, 38%), Asia (226 goats, 20%), Oceania (25 goats, 2%) and Caribbean (16 goats, 1%). This geographical distribution is shown on Figure 1.

Figure 1 here

Figure 1: Geographical distribution of the 1,160 sampled Capra *individuals included in the dataset.*

A unique identification (original ID) was defined for each individual. For resequenced animals and 139 animals published by University of Bern, it starts with 4 letters: the first two correspond to the country of origin (based on ISO 3166-1 alpha-2 codes), the following two define species (CH for *C. hircus*, CA for *C. aegagrus* etc…). Letters 5 to 7 indicate the breed (UNK for unknown) followed by a number that identifies each individual specifically (the meaning of the breed and country codes are explained in Supplementary Table S1). For NextGen data, the sample alias was used as the original ID whereas we used the run accession number for other public data to avoid any duplicate. Concerning the public data, the original ID corresponds to the run accession number.

We also proposed a working name for each individual to facilitate the analysis interpretation for unknown-breed animals. This identifier is composed in the same way as the original IDs presented

above, except that the breed code is followed by the 2 letters of the country before the 4-digit number that identifies each individual.

For public data, information was extracted from BioSample (NCBI, https://www.ncbi.nlm.nih.gov/) in order to determine the sex, breed and geographic origin of each animal. For the VarGoats samples, information was collected by VarGoats collaborators. All individual details are provided in Supplementary Table S2.

As 50k genotypes were available for AdaptMap individuals, we wanted to make sure that no sample had been mixed up. Among available 50k genotypes, 457 were retrieved in the framework of the AdaptMap project which collected samples of 4,653 animals across 148 populations and 35 countries (Stella et al., 2018). Concordance rates (CR) between sequence variants and 50k genotypes of these 457 individuals were checked. Among the 46,654 SNPs of AdaptMap genotypes, 44,691 were found in the vcf files. For each individual the genotype concordance rate was calculated for the SNPs with available genotypes. As shown on Figure 2, the CR is lower for low coverage sequences. For 14 individuals, albeit decent sequencing depth (between 7.2 and 23.6), the CR was lower than 70%, thus indicating a technical problem resulting in a lack of correspondence between the genotyped and sequenced samples (Figure 2, A.).

Figure 2 here

Figure 2: AdaptMap samples analysis: A.concordance rates of the 457 AdaptMap individuals calculated on the 46,654 common SNPs between sequence data and 50k genotypes ; B. PCA to verify the breed of outlier animals.

To check the correctness of the breed indicated in the ID of these animals we performed a Principal Component Analysis (PCA) with PLINK version 1.9 [19] taking into account all the animals belonging to the breeds with problematic samples. The genetic data come from the SNPs contained on autosomes in the VCF file. First of all we removed markers with more than 5% of missing data. This filtering step

yielded 1,890,194 SNPs which were pruned, using the indep-pairwise function on PLINK. Each SNP that had a squared correlation value (R^2) greater than 0.1 with another SNP of a 50-SNPs sliding window moving by 10 SNPs each time (--indep-pairwise 50 10 0.1) was removed. This step reduced their number to 667,949 which were used for the analysis.

Twelve of the fourteen animals clustered with their breed counterparts (Figure 2, B.), while the breed assignment of 2 individuals (AUCH-BOE-0038 and ITCH-CCG-0014) could not be confirmed.

AdaptMap IDs of the 11 animals with missing genotypes are italicized. The 14 low concordance rate individuals are written in bold. The 2 outliers with unconfirmed breed status were labeled as UNK. These 25 problematic individuals have been reported using an asterisk in Supplementary Table S2.

As recommended for any dataset, we advise to perform a global analysis, like a PCA or a structural analysis on the dataset or subset before deeper and specific studies to identify any outlier.

Breed information

The VarGoats dataset encompasses a total of 8 wild species (BEZ, CAU, CPY, CYL, FAL, IBX, NUB, SIB) and 125 breeds belonging to the *C. hircus* species (Supplementary Table S1). They are distributed as follows: 46 in Africa (35%), 40 in Europe (30%), 34 in Asia (25%), 4 in Oceania (3%) and 1 in Caribbean (1%).

The geographical distribution of the breeds was determined based on bibliographical research (www.fao.org; www.racesdefrance.fr; eng.agraria.org/goat.htm; etc.) which allowed us to define more precisely the provenance of each one of them (barycentre of province or geographic region). In the absence of precise information, GPS coordinates of the barycentre of the country of origin were assigned to the sample. Their locations are represented for each continent in Figures 3, 4, 5, except for Creole which is the only one from Caribbean (West Indies, to be more precise). The dataset includes cosmopolitan breeds such as Alpine, Boer or Saanen as well as local breeds that are only present in specific regions of the world.

Figure 3 here

Figure 3: Geographic distribution of European breeds represented by 3 letters corresponding to the breed code (Supplementary Table S1). If a breed is present in several countries, the breed code is followed by the country code (2 letters). Each combination of color and symbol corresponds to domestic goats in a single country, the wild goats are identified with a specific colour and symbol.

Figure 4 here

Figure 4: Geographic distribution of African breeds represented by 3 letters corresponding to the breed code (Supplementary Table S1). If a breed is present in several countries, the breed code is followed by the country code (2 letters). Each combination of color and symbol corresponds to domestic goats in a single country.

Figure 5 here

Figure 5: Geographic distribution of Asian and Oceanian breeds represented by 3 letters corresponding to the breed code (Supplementary Table S1). If a breed is present in several countries, the breed code is followed by the country code (2 letters). Each combination of color and symbol corresponds to domestic goats in a single country, the wild goats are identified with a specific colour and symbol.

Methods

Library construction and sequencing

The library preparation protocol was chosen on the basis of the DNA extraction yield. When available, 250 ng of genomic DNA were sonicated using the E210 Covaris instrument (Covaris, Inc., USA) and the NEBNext DNA Modules Products (New England Biolabs, MA, USA) were used for end-repair, 3'-adenylation and ligation of NextFlex DNA barcodes (Bioo Scientific Corporation). After two consecutive 1x AMPure XP clean-ups, the ligated fragments were amplified by 12 PCR cycles by using the Kapa Hifi Hotstart NGS library Amplification kit (Kapa Biosystems, Wilmington, MA, USA), followed by 0.6x AMPure XP purification. When the nucleic acids extraction yielded low DNA quantities, 10-50 ng of

genomic DNA were sonicated. Fragments were end-repaired, 3'-adenylated and NEXTflex DNA barcoded adapters were added by using NEBNext Ultra II DNA Library prep kit for Illumina (New England Biolabs). After two consecutive 1x AMPure clean-ups, the ligated products were PCR-amplified with NEBNext® Ultra II Q5 Master Mix included in the kit, followed by 0.8x AMPure XP purification.

All libraries were subjected to size profile analysis, with an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) as well as to qPCR quantification (MxPro, Agilent Technologies, USA). Libraries were sequenced using 151 base-length read chemistry in a paired-end flow cell on an Illumina HiSeq 4000 sequencer (Illumina, USA). On average, 120 million paired-end reads were obtained for each sample after clean up (12X expected genomic coverage).

Libraries for the African animals were prepared by Edinburgh Genomics using the TruSeq Nano DNA High Throughput library preparation kit (Illumina, USA). The following preparation protocol was applied: 1 µg of genomic DNA was sheared to fragments of 450bp mean size using a Covaris LE220 focused-ultrasonicator (Covaris, Inc., USA). DNA fragments were then blunt ended, A-tailed, size selected and adapters were ligated into fragment ends according to the Illumina TruSeq PCR-free library preparation kit protocol. Insert size of the libraries was evaluated using a PerkinElmer LapChip GX Touch with an HT DNA 1k/12K/HI SENS LabChip and HT DNA HI SENS Reagent Kit (PerkinElmer, Inc., MA, USA). Final library concentration was calculated by qPCR using a Roche LightCycler 480 (Roche Molecular Systems, Inc., Switzerland) and a Kapa Illumina Library Quantification kit and Standards. Then libraries were normalized to a loading concentration of 150 nM. All the library processing steps were carried out on Hamilton MicroLab STAR liquid handling robots coupled to BaseSpace Clarity LIMS X Edition. Libraries were loaded into a HiSeq X Flow cell v2.5 and clustered using an Illumina cBot2. Clustered flow cells were sequenced at 15X coverage using a HiSeq X Ten Reagent kit v2.5.

An Illumina filter was applied to remove the unreliable data from the analysis. Raw data were filtered to remove clusters with excessive intensity in bases other than the called one. Adapters and primers were removed from the whole read and low quality nucleotides were trimmed from both ends (quality

value lower than 20). Sequences between the second unknown nucleotide (N) and the end of the read were also removed. Reads shorter than 30 nucleotides after trimming were discarded. Finally, the reads and their mates that mapped onto run quality control sequences (PhiX genome) were removed. These trimming steps were achieved using an in-house software based on the FastX package (http://www.genoscope.cns.fr/externe/fastxtend/).

Data generation and preparation

Read alignment and variant calling

Each sample was processed using a pipeline based on the domain reference tools and the Genome Analysis Toolkit (GATK) best practices: BWA version 0.7.15 for alignment [20], SAMtools version 1.6 for handling SAM/BAM formats and calling variants [21], Picard tools version 2.1.1 for labelling duplicated reads (http://broadinstitute.github.io/picard), as well as GATK version 3.6 for Insertion/Deletion (INDEL) realignment, base recalibration and calling variants [22], BCFtools version 1.6 for handling VCF/BCF formats, Freebayes version 1.1.0 for calling variants [23], and snpEff version 4.3t for VCF annotation [24].

Reads were mapped to the latest ARS1 genome version (Genbank accession GCA_001704415.1) of the *C. hircus* species [25] using BWA-MEM software with default parameters except for "-t 14 -M" and "-R" to add read groups. The SAM output files were converted to sorted BAM using SAMtools.

Pre-processing steps (marking duplicates, INDEL-based realignment and base quality score recalibration - BQSR) were done using Picard-MarkDuplicates and GATK. The "known variants" file needed for the BQSR step was computed on a subset of 13 samples (see details in the next paragraph). Variants fulfilling the following conditions were included in the "known variants" file: (1) presenting at least 6 genotypes harbouring an alternative allele ("snpSift filter 'countVariant()>6'") (2) being called by both Freebayes and GATK-HaplotypeCaller.

Variant calling per sample was done with GATK-HaplotypeCaller in ERC mode with a minimum read mapping quality of 30 (this is required to consider a read as valid) and a minimum phred-scaled

confidence threshold of 30 ("-stand_call_conf 30.0 -mmq 30 -ERC GVCF -variant_index_type LINEAR -variant_index_parameter 128000").

Due to the large number of samples, GVCF files were combined (CombineGVCFs) before the joint genotyping step (GenotypeGVCFs) to produce the raw VCF files by chromosome/scaffold.

Filtering process

A Variant Quality Score Recalibration (VQSR) step was performed on the raw VCF files. In order to set up training resource sets for VQSR calibration, 13 goats (AUCH-CAS-0038, BFCH-DJA-0012, CHCH-BOE-0229, ESCH-PAL-0008, ESCH-RAS-0011, ETCH-ABR-0036, FICH-LNR-0122, FRCH-ALP-0030, FRCH-CRE-0014, FRCH-SAA-0032, ITCH-GGT-0026, MZCH-PAF-0003 and ZACH-ANG-0374), representing 13 breeds were chosen out of the 248 animals sequenced in the first batch. They belong to 11 out of the 15 gene pools determined by Colli et al. [15] with the software Admixture [26]. We added one individual from an inbred breed (Palmera) and an animal of the Creole breed (different from the Creole individuals genotyped in the ADAPTmap dataset), as a representative of the American gene pool. Two true sites training resources were built. The first one, corresponding to the highest quality calls, used the variants consistently identified with GATK, Mpileup and Freebayes ("known=false,training=true,truth=true,prior=15.0"). The second one used the 60,000 SNPs selected by Tosser-Klopp et al. [4] to generate the set of polymorphisms included in the Goat SNP50 BeadChip (Illumina) ("known=false,training=true,truth=true,prior=12.0"). The non-true sites training resource was built using the variants exclusively called by GATK ("known=false,training=true,truth=false,prior=10.0").

We only retained biallelic SNPs with a GATK quality score (QUAL) over 100 and with at least two individuals carrying the alternative allele.

Filtering resulted in a high confidence set of 74,274,427 SNPs and 13,607,850 INDELs (Additional Table 1).

Main features of the dataset

The GCF_001704415.1_ARS1_genomic.gff annotation file was used to annotate the variants using SnpEff. Additional Table 1 summarizes the variant types found on each chromosome.

Additional Table 1 here

Sex assignment

As the sex was not determined a priori for each sequenced animal, 100 SNPs mapping on to Y-chromosome contigs were selected (NW_017189563.1, NW_017189610.1, NW_017189618.1, NW_017189628.1, NW_017189685.1, NW_017189696.1, NW_017189885.1, NW_017189985.1, NW_017190040.1, NW_017190154.1, NW_017195709.1). An unknown genotype for these 100 Y-chromosome variants was used as evidence that the sequenced individual was a female. Therefore, for each variant, the number of unknown genotypes was counted for 69 Alpine and Saanen goats from the VarGoats dataset for which the sex was known (i.e., 65 males and 4 females). Among the 100 variants, 17 were retained as they successfully predicted the sex of all these individuals. For assigning the sex of other sampled goats, an individual was considered to be a male when it showed more than genotypes for more than half of 17 aforementioned variants (i.e. 8 genotypes).

Genetic diversity

A subset of 100,000 SNPs was randomly extracted, using the thin-count function from PLINK, from the 667,949 SNPs described for the PCA analysis (individual selection section). Starting from this reduced dataset, a matrix of between-population Reynolds distances was calculated with hapFLK v.1.3.0 [27,28]

and used to construct a Neighbor-Joining tree (Figure 6). The wild species closest to domestic goats (*C. aegagrus*) was used as an outgroup to root the tree.

Figure 6 here

Figure 6: Neighbor-joining tree representing the genetic diversity of domestic goat populations analysed in the context of the VarGoats project.

In the Neighbor-joining tree displayed in Figure 6, goats grouped according to their ancestral geographic provenance (Africa, Europe, Asia or Middle East). This outcome was consistent also for the Oceanian breeds, since Boer goats from Australia (BOE_AU) and New Zealand (BOE_NZ) clustered with other populations of the Boer breed sampled in Africa. Similarly, Australian Cashmere (CAS_AU) grouped with Angora goats due to their common Middle Eastern origin. A similar clustering pattern based on the ancestral geographic origin could be also observed in transboundary breeds sampled in different countries (i.e., Alpine, Boer and Saanen).

The first branching among domestic goats separated the Asian breeds from the remaining populations. This branch carried the most basal cluster which was mainly composed of breeds originating in South-Western Asia (Iranian and Pakistani goats) and thus geographically close to the center of domestication [1]. The second branching included a cluster composed by long-haired breeds, then, two additional African and European clusters emerged in less basal positions. In each continental group, geographically coherent sub-clusters were clearly discernible (e.g. Northern and Southern Europe, North-Western Africa, Eastern Africa and Madagascar).

Concerning the two animals with an ambiguous breed status: one of them (ex AUCH-BOE-0038) grouped with long-haired goats while the other one (ex ITCH-CCG-0014) did so with Pakistani goats.

Re-use potential

The analyses planned in the framework of the VarGoat project are expected to provide a deeper understanding of the evolutionary history of domestic goats and their wild relatives.

The VarGoats data set will be updated with publicly available sequences plus a last batch (60 sequenced at the Genoscope + 38 public sequences extracted on the 2020/02/14), by the end of 2020, containing new sequences generated by our Consortium.

Availability and requirements

The VarGoats dataset is publicly available in the European Nucleotide Archive (ENA) as project number PRJEB37507, which includes fastq and sample description data for 266 animals under accession PRJEB31857, 337 animals under accession PRJEB37122, 29 animals under accession PRJEB37276 and 20 animals under PRJEB37208. The 291 additional sequences used in this article were retrieved from public databases and 217 from NextGen Consortium projects (PRJEB3134, PRJEB3135, PRJEB4371, PRJEB5166, PRJEB3136 and PRJEB5900 studies). Individual accession numbers are listed in Supplementary Table S2.

Use of the data is regulated by a data sharing agreement which is available here: http://www.goatgenome.org/vargoats_agreement.html.

This agreement states that it is mandatory to contact the VarGoats steering committee to discuss the utilization and inclusion of data generated by the VarGoats consortium in any present or future publication. No publications can be generated from the Vargoats data set until the main papers derived from this project are published in scientific journals.

VarGoats project: sample providers and affiliations

James Kijas, CSIRO, Australia

Bernt Guldbrandtsen, Aarhus University, Denmark

Juha Kantanen, Luke, Finland

Dylan Duby, Museum National d'Histoire Naturelle, France

Pierre Martin, Capgenes, France

Coralie Danchin, Delphine Duclos, Institut de l'Elevage, France

Daniel Allain, Rémy Arquet, Nathalie Mandonnet, Michel Naves, Isabelle Palhière, Rachel Rupp, INRAE, France and CABRICOOP breeders

François Pompanon, LECA, France

Hamid R. Rezaei, Gorgan University of Agricultural Sciences and Natural Resources, Iran

Sean Carolan and Maeve Foran, Old Irish Goats Society, Ireland

Alessandra Stella, IBBA-CNR, Italy

Paolo Ajmone-Marsan, Licia Colli, Alessandra Crisà, Donata Marletta and Paola Crepaldi, Italian Goat Consortium, Italy

Michele Ottino, Parco Nazionale del Gran Paradiso, Italy

Ettore Randi, ISPRA Istituto Superiore per la Protezione e la Ricerca Ambientale, Italy

Badr Benjelloun, INRA Maroc, Morocco

Hans Lenstra, University of Utrecht, The Netherlands

Muhammad Moaeen-ud-Din and Jim Reecy, PMAS-Arid Agriculture University, Pakistan

Felix Goyache and Isabel Alvarez, Área de Genética y Reproducción Animal del Serida, Spain

Marcel Amills and Armand Sànchez, Centre for Research in Agricultural Genomics (CRAG), Spain

Juan Capote, Instituto Canario de Investigaciones Agrarias (ICIA), Spain

Jordi Jordana, Universitat Autònoma de Barcelona, Spain

Agueda Pons, Serveis de Millora Agrària i Pesquera (SEMILLA), Illes Balears, Spain

Amparo Martínez and Antonio Molina, University of Córdoba, Spain

Benjamin Rosen, USDA/ARS, United States of America

Carina Visser, Faculty of Natural and Agricultural Sciences, South Africa

Cord Drögemüller, University of Bern, Switzerland

Clet Wandui Masiga, Tropical Institute of Development Innovations (TRIDI), Uganda

Denis Fidalis Mujibi, International Livestock Research Institute (ILRI), Nairobi Kenya

Hassan Ally Mruttu, Ministry of Livestock and Fisheries Development, Tanzania

Timothy Gondwe, Department of Animal Science, Lilongwe University of Agriculture and Natural Resources, Malawi

Joseph Sikosana, Department of Research and Specialist Services, Division of Livestock Research, Zimbabwe

Maria Da Gloria Taela, Animal Production Institute, Ministry of Agriculture, Maputo, Mozambique.

Oyekan Nash, National Biotechnology Development Agency, Nigeria

Additional files

Additional Table 1:

Title: INDELs distribution and annotation of SNPs identified when sequences were aligned to the ARS1 reference genome and using the GCF_001704415.1_ARS1_genomic.gff annotation file.

Format: .xlsx file

Description: Number of INDELs and SNPs identified per chromosome (CHI) and annotation information for SNPs identified

Supplementary Table S1:

Title: Number of individuals per breed and country of origin

Format: .xlsx file

Description: Distribution of sequenced individuals per breed and abbreviations explanation

Supplementary Table S2:

Title: Detailed information for each sequenced individual

Format: .xlsx file

Description: Description of each individual (species, breed, country of origin, localization, sex, sample provider and details about its sequence)

List of abbreviations

BQSR = Base Quality Score Recalibration

CR = Concordance Rate

GATK = Genome Analysis ToolKit

INDEL = Insertion/Deletion

PCA = Principal Component Analysis

SNP = Single Nucleotide Polymorphism

VQSR = Variant Quality Score Recalibration

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

LD, ET, GTK performed analyses and wrote the paper

PB performed the sequence analysis, generated the VCF files, compared genotyping and sequencing data.

LC, IP, RR, CD chose the relevant samples to represent international genetic diversity.

JS performed DNA extraction and managed the DNA samples for Génoscope.

MA and LC contributed to the design of the project and the correction of the paper.

AA, CO and SE managed whole genome sequence production and bioinformatic quality control in Genoscope.

MDC performed additional checks on the variation level of post-QC data.

FP contributed to the correction of the paper, and co-supervised LD PhD work with GTK.

CWM contributed to the design of AGIN, lead sampling in Uganda, Tanzania, Malawi, Mozambique and Zimbabwe and participated in improving and advancing the paper.

EC received the DNA samples from African animals from BR, prepared and submitted them to Edinburgh Genomics for sequencing, and coordinated the research grant under which they were sequenced. MS oversaw management and upload to the ENA of the African dataset. So did PB, SE and GTK for the other datasets.

MA, EC, LC, PC, TF, FP, BR, AS and CVT contributed to the design and the management of the project.

GTK coordinated the project.

Acknowledgements

First of all, we acknowledge the sample providers of VarGoats project (listed above).

We acknowledge the work of Génoscope, USDA and Edinburgh Genomics, University of Edinburgh sequencing platforms that generated the sequencing data for AGIN and VarGoats projects and the GeT-Plage platform for generating data within CAPRISNP and ACTIVEGOAT projects.

We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees, the CTIG (Centre de Traitement de l'Information Génétique) of INRAE Jouy-en-Josas and the TGCC (Très Grand Centre de calcul) of CEA (Commissariat à l'énergie atomique) Bruyères-le-Châtel for providing computing resources.

We thank Cédric Cabau (INRAE) for setting up and maintaining the VarGoats website.

We thank Marina Naval Sanchez (Institute for Molecular Bioscience, the University of Queensland, Australia) and Miguel Perez-Enciso (ICREA, Barcelona, Spain and Centre for Research in Agricultural Genomics, CRAG, Bellaterra, Spain) for sharing scripts and participating in helpful discussions about variant calling and filtering.

We thank Florent Woloszyn (INRAE) for DNA extraction of some French DNA samples and Erwan Quéméré (INRAE) for providing *Capra pyrenaica* samples.

We thank French local breed associations.

We thank the Cabricoop breeders, and Rémy Arquet in charge of the INRAE flock, who provided the animals selected for sampling the Creole goat population. We are grateful to Nathalie Mandonnet (INRAE) who identified the animals representative of the variability of the Creole goat population and Michel Naves (INRAE), scientific animator of the biological resource centre CARARE, who supplied the study with the blood samples.

We thank Delphine Duclos (Institut de l'Elevage) who identified the animals representative of the variability of the Savoie breed.

We thank Dr Stefano Frattini (UNIMI, Italy) for Italian DNA samples preparation.

We thank Oyekan Nash, National Biotechnology Development Agency, Nigeria for DNA extraction of samples from Uganda, Kenya, Tanzania, Malawi, Zimbabwe and Mozambique.

We are grateful to Derek M Bickhart (USDA), Christophe Klopp (INRAE), Ezequiel Nicolazzi (former at PTP), Tad Sonstegard (former at USDA) and George R Wiggans (USDA) who helped writing the France Génomique proposal in 2016.

We thank Professors David Hume and Mick Watson for acquiring the Data and Resources research grant from the Biotechnology and Biological Sciences Research Council from which the sequencing of the African goats was funded and Professor Appolinaire Djikeng for coordination of collaborative networks through the Centre for Tropical Livestock Genetics and Health.

We thank Jeffrey Silverstein, Tad Sonstegard, Curt Van Tassell, Jennifer Woodward-Greene, Max Rothschild, Heather Huson, Brian Sayre, and Hans Soelkner for AGIN design and sampling plan. We thank the teams that participating in sampling in African countries as detailed below: Getinet Mekuriaw Tarekegn (Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences & Department of Animal Production and Technology, Bahir Dar University, Bahir Dar), Aynalem Haile, (International Center for Agricultural Research in Dry Areas, Addis Ababa), Tadelle Dessie (International Livestock Research Institute (ILRI), Addis Ababa), Tesfaye Getachew (International Center for Agricultural Research in Dry Areas, Addis Ababa), Solomon Abegaz (Depa rtment of Animal Production, Woldia University, Woldia) and Barbara Rischkowsky (ICARDA) for Ethiopia ; Denis Fidalis Mujibi (Usomi Limited), Kihara Absolomon and Moses Ogugo (ILRI, Nairobi) for Kenya ; Maminiaina Olivier Fridolin (Département de Recherches Zootechniques et Vétérinaires (FOFIFA-DRZV) du Centre National de Recherche Appliquée au Développement Rural) for Madagascar ; Timothy Gondwe (Animal Breeding, Department of Animal Science, Lilongwe University of Agriculture and Natural Resources, Lilongwe), Wilson Nandolo, (Animal Breeding and Genetics, Department of Animal Science, Lilongwe University of Agriculture and Natural Resources, Lilongwe), Winchester Mvula, Thomson Sanudi, Lchai Mabanda (Department of Animal Science, Bunda College of Agriculture, LUANAR, Lilongwe), for Malawi ; Maria Da Gloria Taela, Evaristo Vasco, Alaento Domingos, Franciso (Animal Production Institute, Ministry of Agriculture, Maputo) , for Mozambique ; Ally Mruttu, Pius Masanja Paul, Julius Budodi and Muhidin H. Shemashinde (Livestock Research, AnGR, Ministry of Livestock and Fisheries Development), for Tanzania ; Henry Mulindwa (National Livestock Resources Research Institute (NALIRRI)), Brian Babigumira, Christopher Mukasa (National Animal Genetic Resources Centre (NAGRC)) and Noah Sabunyo (Tropical Institute of Development Innovations (TRIDI)) for Uganda ; Joseph Sikosana, Milton Makumbe and Josepaht Mukwena (Department of Research and Specialist Services, Division of Livestock Research) for Zimbabwe.

Funding

We are grateful to France Génomique "Call for high impact projects" because of selecting our project and providing us the resources to sequence 400 goats.

We would like to mention that APIS-GENE funded some WGS sequences through ACTIVEGOAT & CAPRISNP projects.

We thank the Occitanie region and the Animal Genetics Division of the French National Institute for Agriculture, Food and Environment (INRAE-GA) for financing E.T PhD.

We thank the Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation for financing L.D.

We thank André Eggen (Illumina) for providing chips to genotype 192 animals.

We thank the Animal Genetics Division of the French National Institute for Agriculture, Food and Environment (INRAE-GA) for funding VarGoats2 grant, which allowed DNA extraction and genotyping of 384 animals and CRB-Anim, Grant Agreement ANR-11-INBS-0003, (https://crb-anim.fr/) for funding French local breeds sampling.

We thank the Italian Goat and Sheep Breeders Association (AssoNaPa) for supporting in sampling.

Whole genome sequencing libraries for the African goats were prepared and sequenced by Edinburgh Genomics and funded via Biotechnology and Biological Sciences Research Council research grant (BBS/OS/GC/000012F) 'Reference genome and population sequencing of African goats' awarded to The Roslin Institute.

USDA-ARS with funding from USAID funded the collection of samples from Uganda, Tanzania, Malawi, Mozambique and Zimbabwe.

EC and MS were partially supported by the Bill & Melinda Gates Foundation and with UK aid from the UK Government's Department for International Development (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh, SRUC (Scotland's Rural College), and the International Livestock Research

Institute. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation nor the UK Government.

References

1. Zeder MA, Hesse B. The Initial Domestication of Goats (*Capra hircus*) in the Zagros Mountains 10,000 Years Ago. Science, New Series. 2000;287:2254–7.

2. Skapetas B, Bampidis V. Goat production in the World: present situation and trends. Livest Res Rural Dev. 2016;28:200.

3. Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, et al. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). Nat Biotechnol. 2013;31:135–41.

4. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al. Design and characterization of a 52K SNP chip for goats. PLoS ONE. 2014;9:e86227.

5. Martin PM, Palhière I, Ricard A, Tosser-Klopp G, Rupp R. Genome Wide Association Study Identifies New Loci Associated with Undesired Coat Color Phenotypes in Saanen Goats. PLoS ONE. 2016;11:e0152426.

6. Martin P, Palhière I, Maroteau C, Bardou P, Canale-Tabet K, Sarry J, et al. A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content. Scientific Reports. Nature Publishing Group; 2017;7:1872.

7. Mucha S, Mrode R, Coffey M, Kizilaslan M, Desire S, Conington J. Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. Journal of Dairy Science. 2018;101:2213–25.

8. Carillier C, Larroque H, Palhière I, Clément V, Rupp R, Robert-Granié C. A first step toward genomic selection in the multi-breed French dairy goat population. Journal of Dairy Science. 2013;96:7294–305

9. Carillier-Jacquin C, Larroque H, Robert-Granié C. Including αs1casein gene information in genomic evaluations of French dairy goats. Genetics Selection Evolution. 2016;48:54.

10. Liu M, Zhou Y, Rosen BD, Van Tassell CP, Stella A, Tosser-Klopp G, et al. Diversity of copy number variation in the worldwide goat population. Heredity. Nature Publishing Group; 2018;122:636–46.

11. Cardoso TF, Amills M, Bertolini F, Rothschild M, Marras G, Boink G, et al. Patterns of homozygosity in insular and continental goat breeds. Genet Sel Evol. 2018;50:56.

12. Bertolini F, Cardoso TF, Marras G, Nicolazzi EL, Rothschild MF, Amills M, et al. Genome-wide patterns of homozygosity provide clues about the population history and adaptation of goats. Genet Sel Evol. 2018;50:59.

13. Bertolini F, Servin B, Talenti A, Rochat E, Kim ES, Oget C, et al. Signatures of selection and environmental adaptation across the goat genome post-domestication. Genet Sel Evol. 2018;50:57.

14. Stella A, Nicolazzi EL, Van Tassell CP, Rothschild MF, Colli L, Rosen BD, et al. AdaptMap: exploring goat diversity and adaptation. Genet Sel Evol. 2018;50:61.

15. Colli L, Milanesi M, Talenti A, Bertolini F, Chen M, Crisà A, et al. Genome-wide SNP profiling of worldwide goat populations reveals strong partitioning of diversity and highlights post-domestication migration routes. Genet Sel Evol. 2018;50:58.

16. Talenti A, Palhière I, Tortereau F, Pagnacco G, Stella A, Nicolazzi EL, et al. Functional SNP panel for parentage assessment and assignment in worldwide goat breeds. Genet Sel Evol. 2018;50:55.

17. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Molecular Biology and Evolution. 2010;27:2534–47.

18. Benjelloun B, Boyer F, Streeter I, Zamani W, Engelen S, Alberti A, et al. An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity. Molecular Ecology Resources. 2019;19:1497–515.

19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81:559–75.

20. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation. Genome Research. 2010;20:1297–303.

23. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:12073907 [q-bio] [Internet]. 2012 [cited 2020 Feb 3]; Available from: http://arxiv.org/abs/1207.3907

24. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118} ; iso-2; iso-3. Fly. 2012;6:80–92.

25. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50.

26. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009;19:1655–64.

27. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. Genetics. 2010;186:241–62.

28. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. Genetics. 2013;193:929–41.

Figure 1

Click here to access/download;Figure;Figure1.pdf

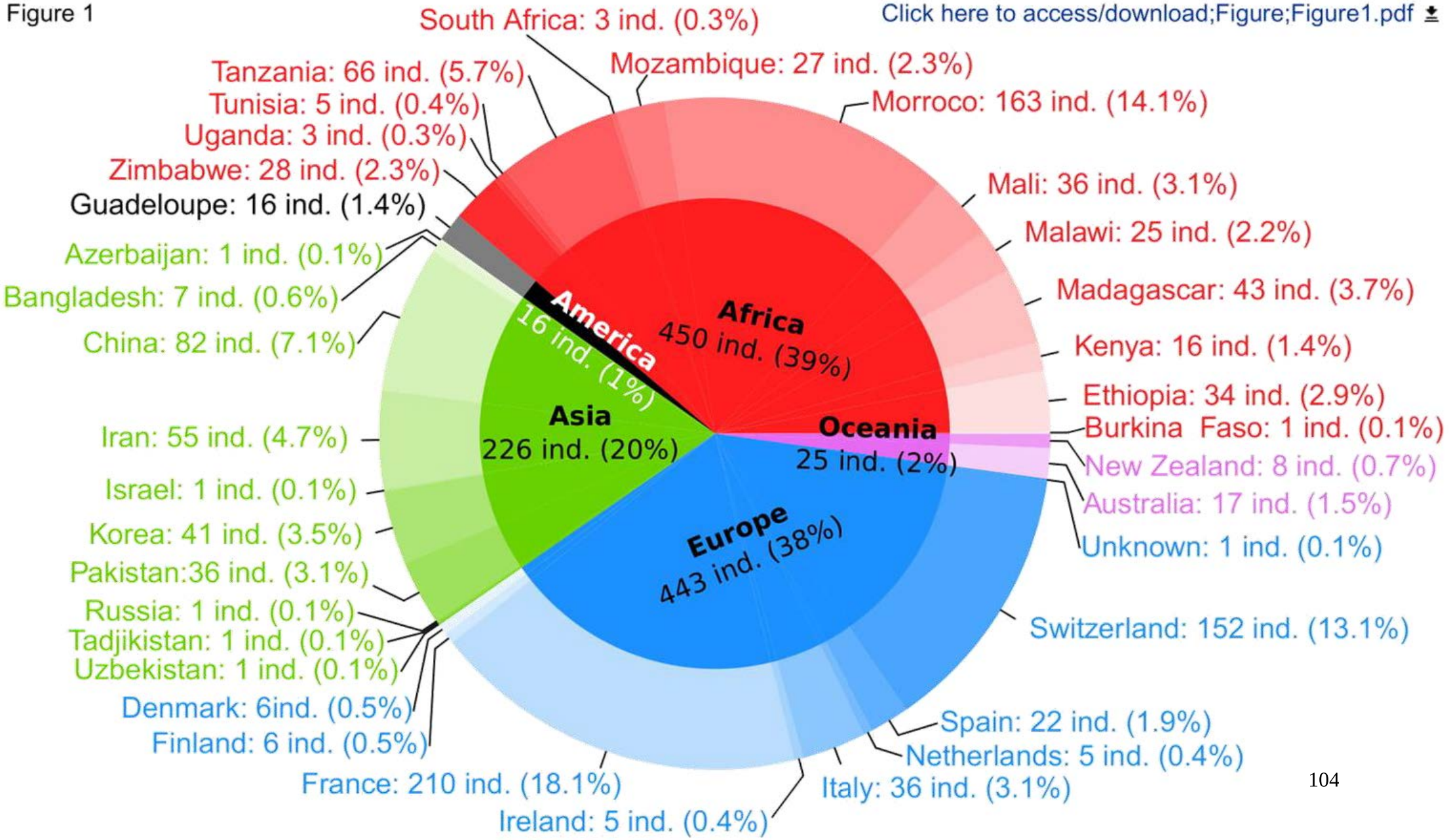


Figure 2

Click here to access/download;Figure;Figure2.pdf

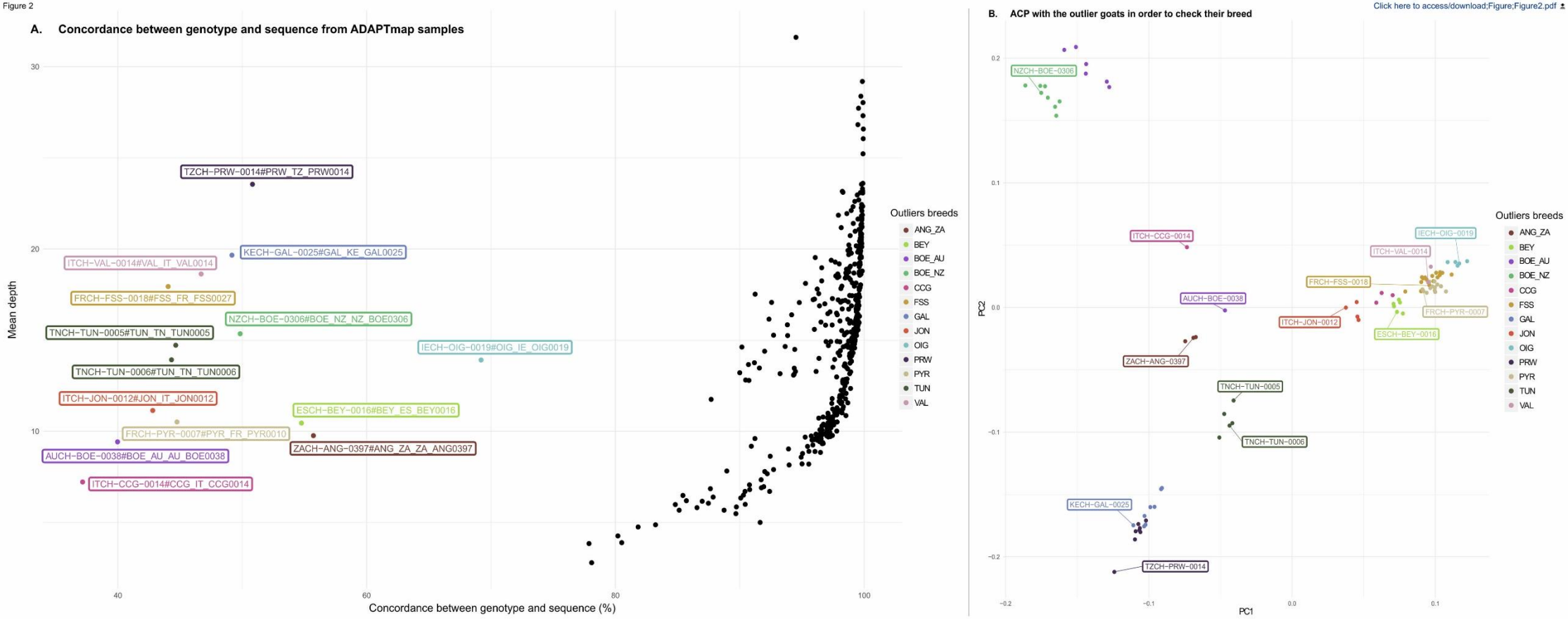


Figure 3

Click here to access/download;Figure;Figure3.pdf

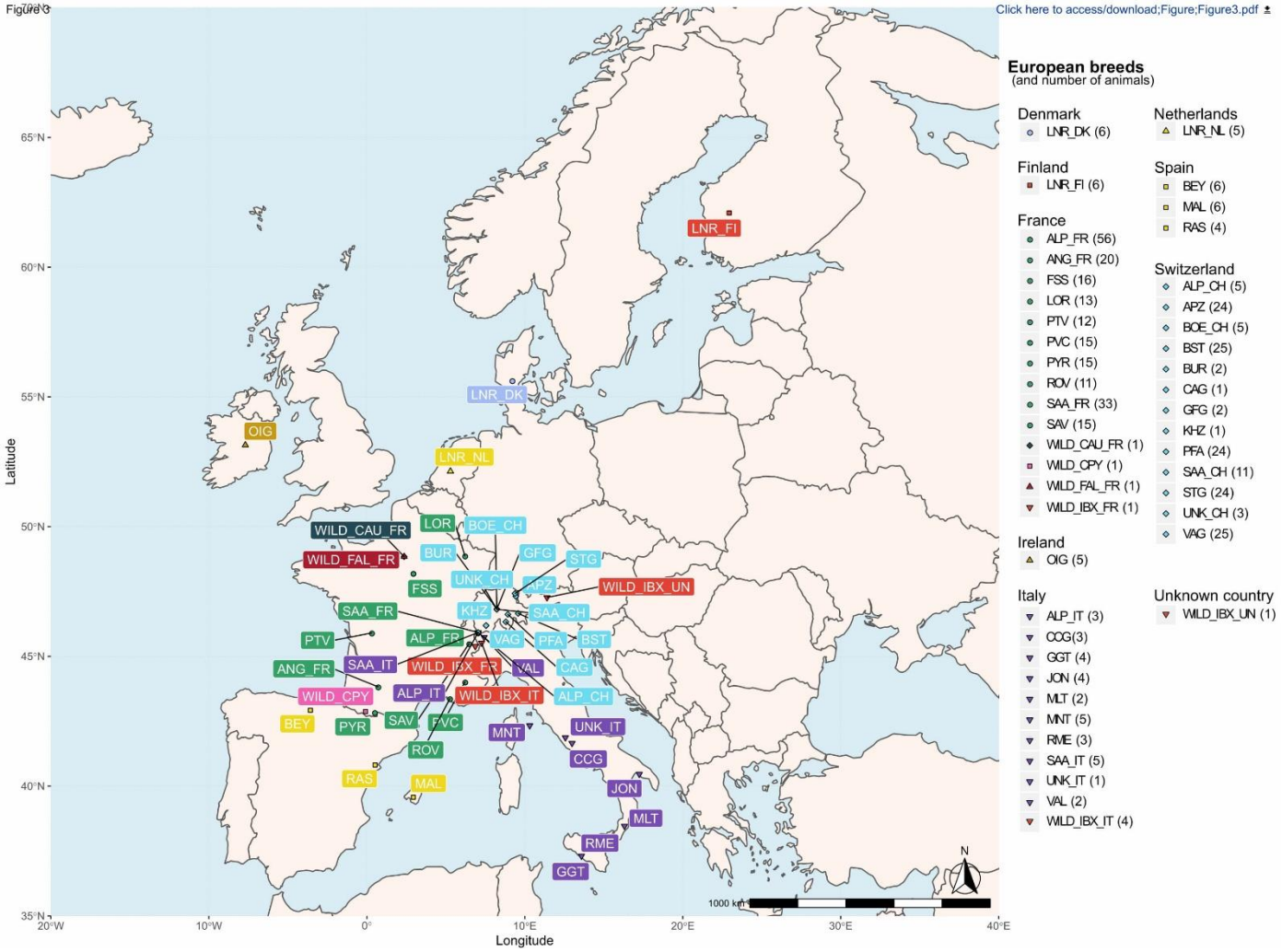


Figure 4

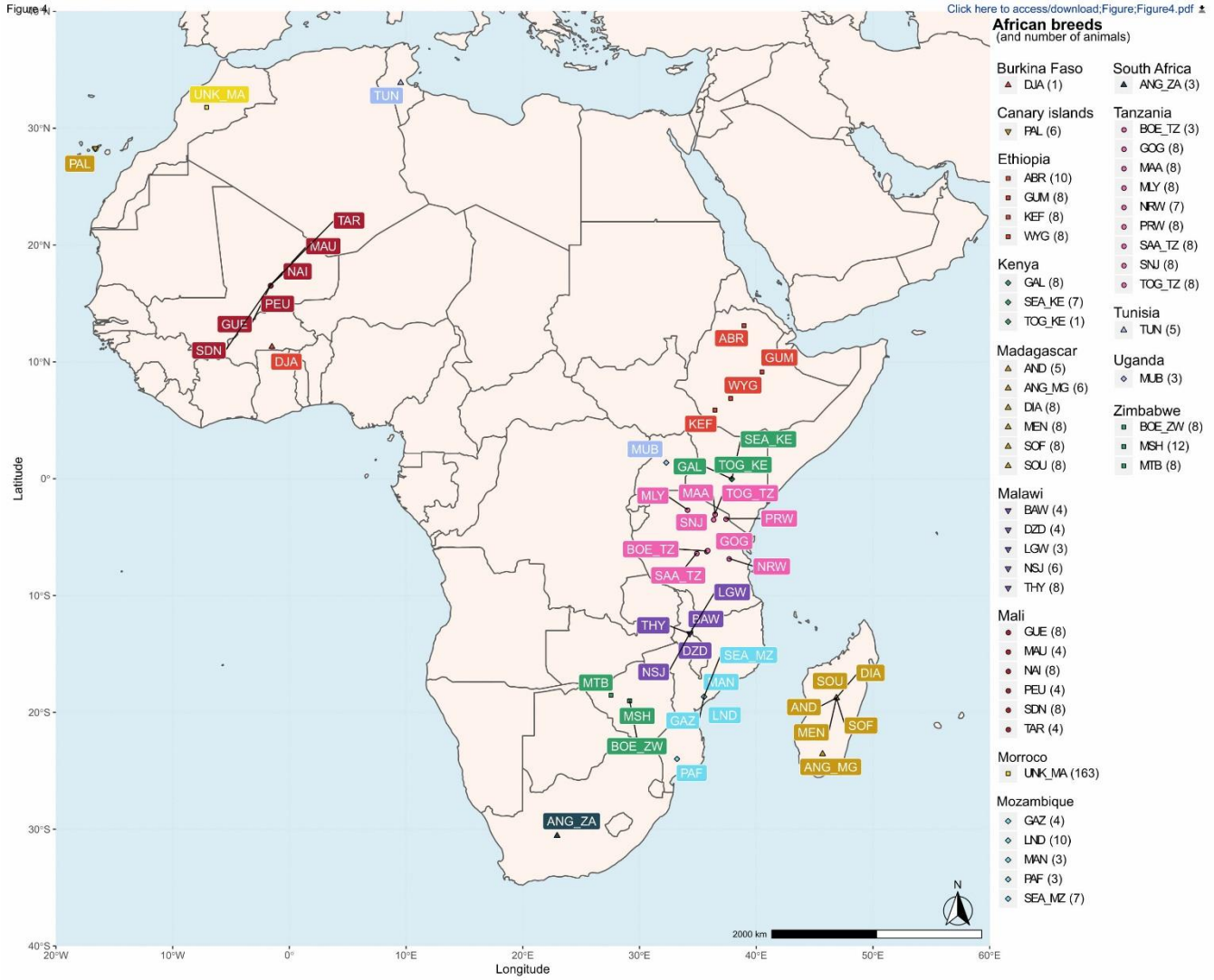

Click here to access/download;Figure;Figure4.pdf
African breeds
(and number of animals)
Burkina Faso
DJA (1)
Canary islands
PAL (6)
Ethiopia
ABR (10)
GUM (8)
KEF (8)
WYG (8)
Kenya
GAL (8)
SEA_KE (7)
TOG_KE (1)
Madagascar
AND (5)
ANG_MG (6)
DIA (8)
MEN (8)
SOF (8)
SOU (8)
Malawi
BAW (4)
DZD (4)
LGW (3)
NSJ (6)
THY (8)
Mali
GUE (8)
MAU (4)
NAI (8)
PEU (4)
SDN (8)
TAR (4)
Morroco
UNK_MA (163)
Mozambique
GAZ (4)
LND (10)
MAN (3)
PAF (3)
SEA_MZ (7)
South Africa
ANG_ZA (3)
Tanzania
BOE_TZ (3)
GOG (8)
MAA (8)
MLY (8)
NRW (7)
PRW (8)
SAA_TZ (8)
SNJ (8)
TOG_TZ (8)
Tunisia
TUN (5)
Uganda
MUB (3)
Zimbabwe
BOE_ZW (8)
MSH (12)
MTB (8)
Latitude
Longitude
2000 km
N

Figure 5

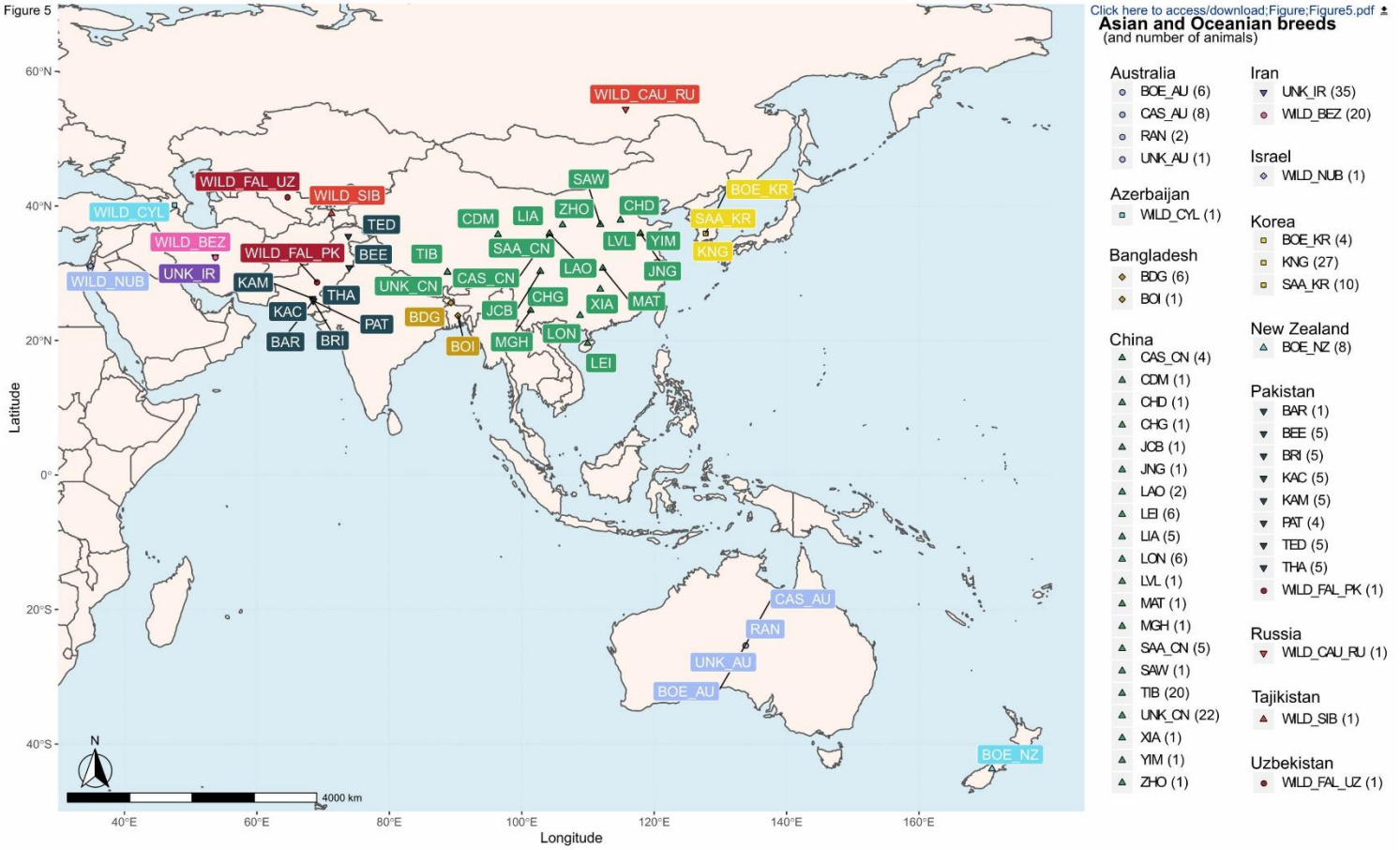
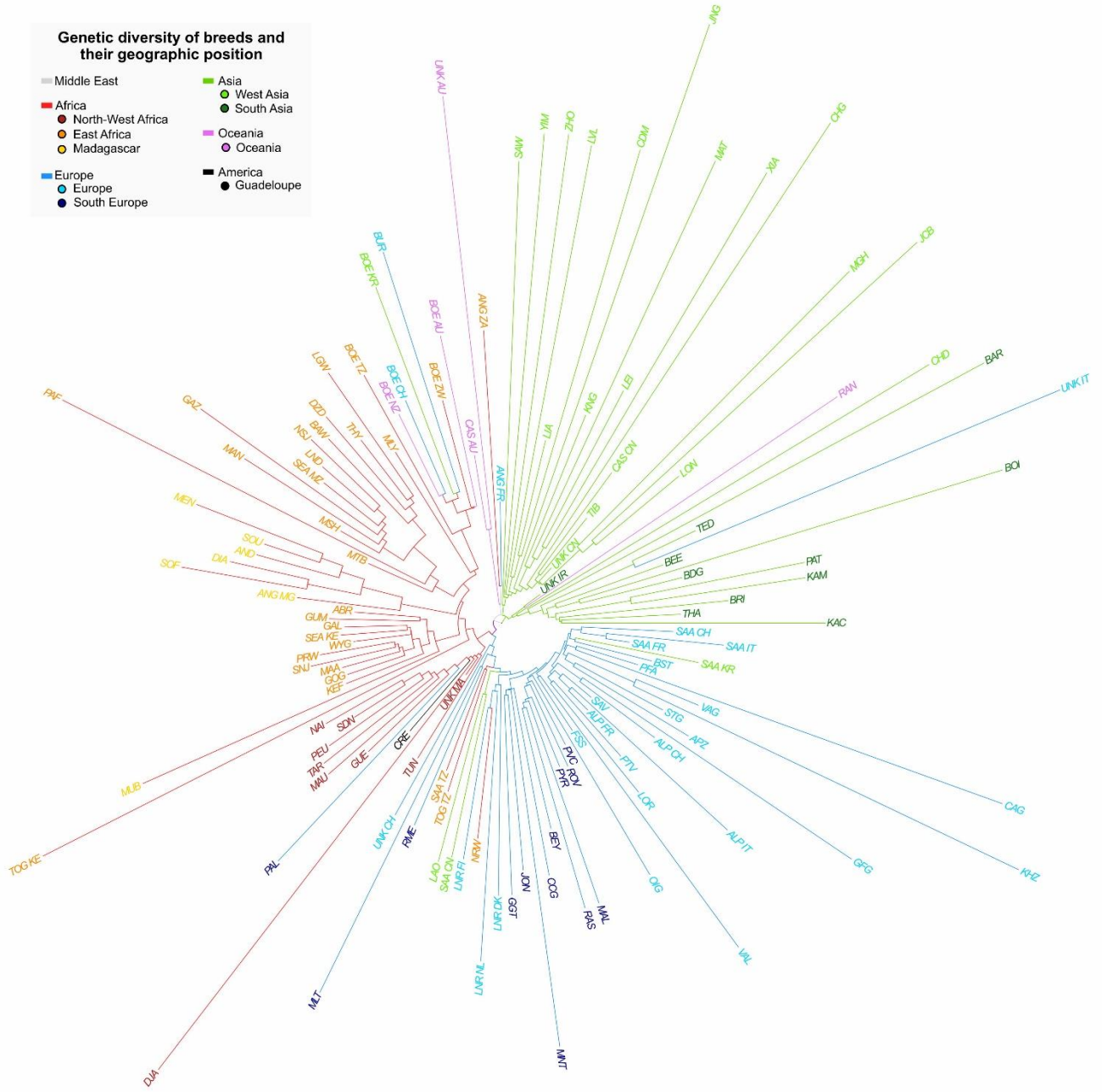

Click here to access/download;Figure;Figure5.pdf
Asian and Oceanian breeds
(and number of animals)
Australia
BOE_AU (6)
CAS_AU (8)
RAN (2)
UNK_AU (1)
Azerbaijan
WILD_CYL (1)
Bangladesh
BDG (6)
BOI (1)
China
CAS_CN (4)
CDM (1)
CHD (1)
CHG (1)
JCB (1)
JNG (1)
LAO (2)
LEI (6)
LIA (5)
LON (6)
LVL (1)
MAT (1)
MGH (1)
SAA_CN (5)
SAW (1)
TIB (20)
UNK_CN (22)
XIA (1)
YIM (1)
ZHO (1)
Iran
UNK_IR (35)
WILD_BEZ (20)
Israel
WILD_NUB (1)
Korea
BOE_KR (4)
KNG (27)
SAA_KR (10)
New Zealand
BOE_NZ (8)
Pakistan
BAR (1)
BEE (5)
BRI (5)
KAC (5)
KAM (5)
PAT (4)
TED (5)
THA (5)
WILD_FAL_PK (1)
Russia
WILD_CAU_RU (1)
Tajikistan
WILD_SIB (1)
Uzbekistan
WILD_FAL_UZ (1)
WILD_CAU_RU
WILD_FAL_UZ
WILD_SIB
WILD_CYL
TED
WILD_BEZ
WILD_FAL_PK
BEE
TIB
UNK_IR
WILD_NUB
KAM
UNK_CN
CAS_CN
THA
KAC
PAT
BAR
BRI
BDG
BOI
JCB
CDM
LIA
SAW
ZHO
CHD
SAA_CN
LVL
YIM
BOE_KR
SAA_KR
KNG
LAO
JNG
CHG
XIA
MAT
MGH
LON
LEI
CAS_AU
RAN
UNK_AU
BOE_AU
BOE_NZ
N
4000 km
60°N
40°N
20°N
0°
20°S
40°S
Latitude
40°E
60°E
80°E
100°E
120°E
140°E
160°E
Longitude

Figure 6

Click here to access/download;Figure;Figure6.pdf



II. Filtrage des données brutes de séquence

Une fois alignées sur le génome de référence, les données de séquences de l'ensemble des individus sont agrégées sous la forme d'un fichier *vcf* (cf Chapitre 1). Ce fichier recense chaque position du génome pour laquelle une variation par rapport au génome de référence a été observée pour un des individus séquencés. Le jeu de données final présenté dans le *data paper* et mis à disposition par le consortium VarGoats a été filtré par Philippe Bardou (INRAE Sigenae) avec quelques filtres basiques et sur la base d'un apprentissage VQSR (Variant Quality Score Recalibration). Le VQSR repose sur du *machine learning*. Plusieurs jeux de données sont fournis à un algorithme pour lui apprendre à distinguer les variants de bonne qualité des autres. Une fois paramétré, l'algorithme est utilisé pour filtrer le jeu de données de l'ensemble des séquences.

Avant l'obtention du jeu de données présenté dans le *data paper* publié par le consortium VarGoats, de nombreux jeux de données intermédiaires ont été mis à disposition par le consortium VarGoats pour permettre aux différents groupes de travail d'amorcer les premières études. Dans le cadre de ma thèse j'ai conduit des premières analyses de qualité dès les premiers jeux de donnés disponibles en 2011, dans l'objectif prioritaire de pouvoir exploiter les données Alpine et Saanen françaises. Ces travaux sont résumés ci-après.

II.1. Evolution du nombre de séquences et des différents jeux de données intermédiaires

Au cours de ma thèse, le nombre de séquences disponibles a évolué et 7 fichiers de séquence ont été produits (Tableau 6). J'ai pu exploiter les 6 premiers dont le nombre de séquences est progressivement passé de 27 à 829. Le filtrage définitif a été adopté sur un jeu de 829 séquences qui contenait déjà les 81 séquences de race Alpine et Saanen, présentes dès l'étape 5 (Tableau 6).

Tableau 6: Différents jeux de données mis à ma disposition au cours de la thèse

NOMBRE TOTAL DE	***NOMBRE DE SEQUENCES ALPINE ET SAANEN***	***VERSION DU GENOME CAPRIN***	***DATE DE MISE A DISPOSITION***

SEQUENCES	*FRANÇAISES*		
27	13 Alpine	CHIR1.0	Octobre 2017
	11 Saanen		
248	30 Alpine	ARS1.0	Mars 2018
	32 Saanen		
62	30 Alpine	ARS1.0	Mars 2018
	32 Saanen		
285	30 Alpine	ARS1.0	Avril 2018
	32 Saanen		
594	44 Alpine	ARS1.0	Août 2018
	37 Saanen		
829	44 Alpine	ARS1.0	Janvier 2019[1]
	37 Saanen		
1160	56[3] Alpine	ARS1.0	Juillet 2019
	33[2] Saanen		

1 dernier fichier traité dans le cadre des analyses d'imputation et d'association

2 les individus séquencés avec une profondeur globale inférieure à 5X et une qualité de génotypage trop faible ont été enlevés du jeu de données

3 ajout de 16 boucs Alpins français fondateurs de lignées divergentes sur la longévité fonctionnelle de l'Unité expérimentale INRAE de Bourges

II.2. Le filtrage des données

Les filtres que j'ai pu tester au cours de ma thèse sont multiples et ont permis d'alimenter la réflexion sur le filtrage du jeu de données qui a ensuite été fourni à l'ensemble des groupes de travail du consortium VarGoats. Cette étude a donc été largement exploratoire et toutes les pistes envisagées ne seront pas présentées.

Le filtrage des données dans cette thèse répond à une double problématique :

- éliminer le plus grand nombre d'erreurs de séquençage et ainsi éviter de propager des erreurs dans les séquences imputées
- conserver suffisamment de variants pour ne pas écarter des mutations intéressantes pour des analyses ultérieures

Nous pouvons catégoriser les filtres appliqués aux données de séquence caprines en deux grands groupes : les filtres généraux et les filtres spécifiques qui ont été développés grâce aux données de génotypages disponibles en Alpine et Saanen françaises. Les deux types

de filtres sont appliqués à l'intégralité du jeu de données (829 séquences). Les étapes du filtrage et l'évolution du nombre de variants sont présentées sur la Figure 21.

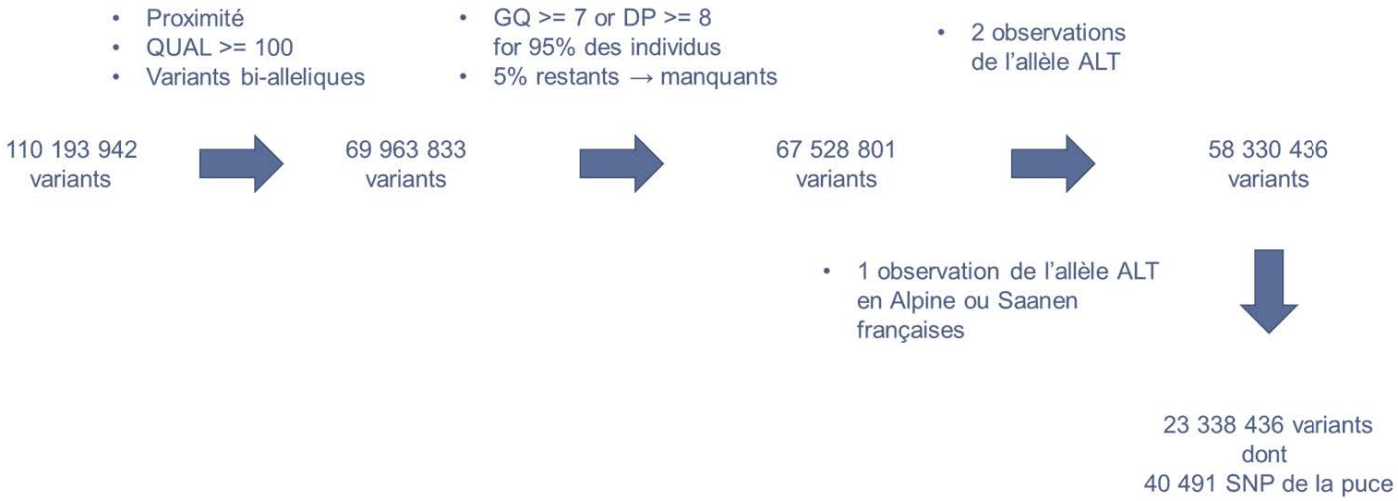


Figure 21: Les étapes du filtrage retenues pour la suite des travaux de thèse

II.2.a. Filtres généraux appliqués à l'ensemble des séquences

Avant toute analyse, quatorze individus séquencés (dont 4 Alpines et 4 Saanen françaises) ont été écartés du jeu de séquences initial. Ils présentaient des profondeurs de séquence très faibles (couverture inférieure à 5X) et se sont avérés être source d'erreurs d'imputation dans la suite des analyses.

Dans la version finale du filtrage des données, les filtres généraux interviennent à plusieurs stades. Ainsi dans un premier temps, le variant de plus grande qualité est conservé (paramètre QUAL estimé par l'étape de recalibration des bases de GATK qui rend compte de la probabilité qu'il y ait plusieurs allèles à cette position) quand deux indels sont trop proches (distance inférieure à 10 pb). De même, quand deux variants SNP ou indels sont distants de moins de 3 pb, le variant de plus faible qualité est éliminé. Ce même type de filtre a été appliqué sur les données de séquence bovines (Daetwyler et al., 2014). Le *calling* de deux variants proches peut être la résultante d'une mauvaise interprétation des lectures par l'algorithme de *calling*. En effet, une insertion ou une délétion produit des lectures plus difficiles à aligner sur le génome de référence qu'un simple SNP qui est par définition ponctuel. Il arrive alors que l'algorithme interprète comme deux variants distincts une insertion/délétion. En parallèle, grâce à l'outil SnpSift (Cingolani et al., 2012), et suite à des

échanges avec Mekki Boussaha (UMR GABI, INRAE), nous avons décidé d'éliminer les variants dont la qualité est inférieure à 100. Cette valeur correspond à un taux d'erreur maximum de 10^{-10}. Ce critère est un peu plus strict que ce qui a pu être implémenté sur les séquences bovines (QUAL ≥ 20) (Daetwyler et al., 2014), toutefois il élimine finalement assez peu de variants. Enfin, les variants présentant plus de deux allèles ont été systématiquement écartés car leur utilisation dans la suite des analyses (imputation, analyses d'association, évaluations génomiques) est complexe. Cette première étape de filtrage est conséquente et élimine plus de 40 millions de variants du jeu de données.

Un second filtre général a été appliqué dans les étapes suivantes. Il écarte tous les variants pour lesquels l'allèle alternatif est observé moins de 2 fois (ALT) sur toutes les séquences. Il faut donc au minimum deux porteurs hétérozygotes ou un homozygote pour l'allèle alternatif d'un variant pour que ce dernier soit conservé. Ce filtre écarte les variants avec des allèles trop rares. Les *Capra* sauvages étaient potentiellement les uniques porteurs d'allèles présentant peu d'intérêt pour la majorité des études conduites sur toutes les données caprines (en dehors de l'étude des traces de sélection).

II.2.b. Filtres conçus à partir de typages disponibles en races Alpines et Saanen

Les travaux de cette thèse se concentrent sur les races Alpine et Saanen françaises. Dans ces deux races, à l'exception de 3 Alpines, nous disposons pour les animaux séquencés de leurs génotypes sur la puce 50k caprine. Il était donc possible de comparer les génotypes obtenus à des positions comparables du génome. La comparaison des génotypes des marqueurs de la puce 50k à ceux équivalents sur la séquence ne peut pas être réalisée directement mais nécessite au préalable d'établir une correspondance. En effet, le brin à partir duquel la sonde de la puce à ADN a été construite et le brin lu sur le génome de référence ne sont pas nécessairement les mêmes. L'information de brin des sondes de la puce nous a permis de permuter les allèles des génotypages 50k pour qu'ils correspondent aux allèles observés sur la séquence. Plus tard, il s'est avéré que nous n'avions dans nos génotypes 50k que des SNP avec des allèles A/G ou A/C pour lesquels l'établissement d'une correspondance ne nécessite pas d'avoir l'information du brin des sondes Illumina (pas d'inversion possible). Ceci nous a permis de ne plus à avoir à faire de tableau de correspondance et ainsi de réduire les temps de calcul nécessaires à la conversion.

Nous avons voulu mettre en regard le résultat de cette comparaison avec des paramètres de qualité locaux définis par GATK. La profondeur (DP) et la qualité de génotype (GQ) d'un individu à une position variante donnée nous sont apparues comme de bons indicateurs pour distinguer les erreurs de séquençage des génotypes fiables. En effet, le pourcentage de variants discordants était plus élevé dans les séquences de faible couverture. A titre d'exemple, lors d'un précédent filtrage sur un jeu de 594 séquences, nous avons comparé la profondeur moyenne des 81 individus de race Alpine et Saanen sur le chromosome 1 et le pourcentage d'erreurs de concordance sur ce chromosome (Figure 22).

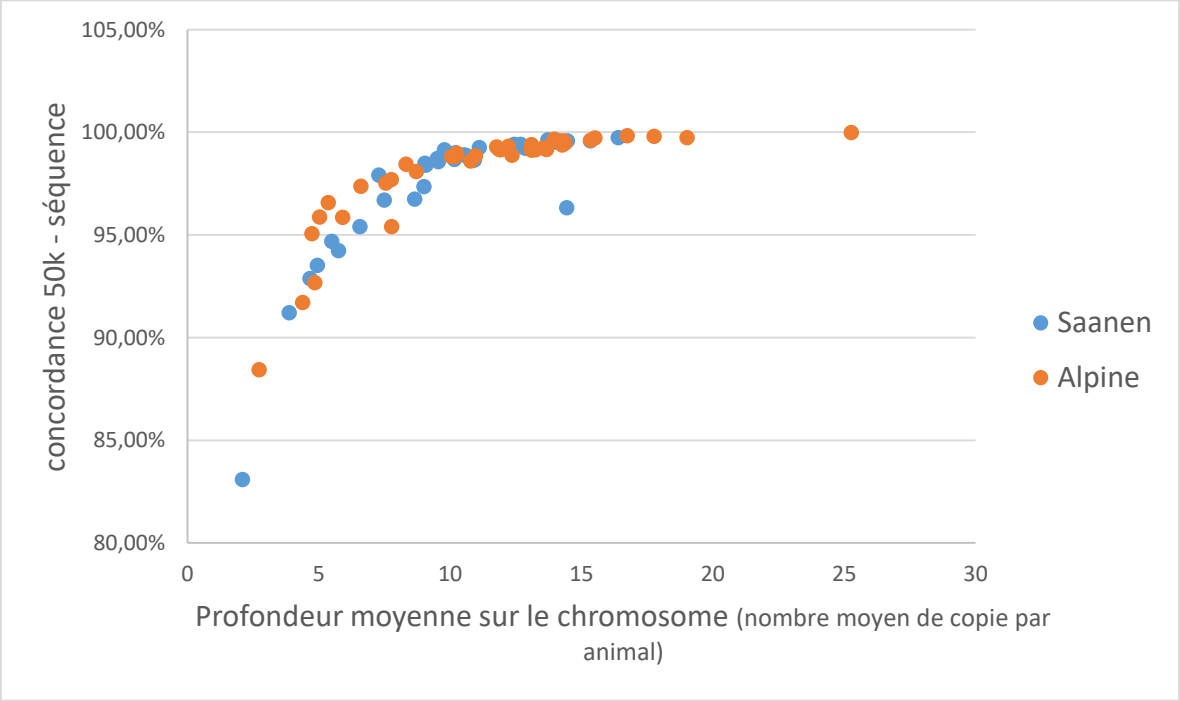


Figure 22: Taux d'erreurs de concordance entre génotypes obtenus à partir de la puce 50k et obtenus à partir de la séquence d'un individu sur le chromosome 1.

De plus, nous avons pu constater que ces paramètres prenaient des valeurs différentes selon que les génotypes des mêmes variants correspondaient entre séquences et génotypes 50k ou non (Tableau 7). Ces valeurs nous ont permis de fixer des seuils généraux ensuite appliqués à l'ensemble des données. Ainsi un variant n'est conservé que si 95% des animaux présentent un DP ≥ 8 ou un GQ ≥ 7. Le filtre définitif est issu d'un compromis. En effet, les deux conditions ne peuvent être appliquées simultanément sans entrainer la perte de tous les variants disponibles. Les variants qui passent ce filtre doivent donc présenter un DP ≥ 8 ou un GQ ≥ 7 pour au minimum 95% des animaux séquencés. Les génotypes sont remplacés par des

génotypes manquants pour les quelques animaux (maximum 5%) qui ne respectent pas la condition.

Tableau 7: Profondeur (DP) et qualité de génotype (GQ) locales en fonction de la concordance avec les génotypes 50k des 41 Alpines et 37 Saanen français séquencés et génotypés

	DP			***GQ***		
	Moyenne	Min	Max	Moyenne	Min	Max
GENOTYPES CONCORDANTS	11,7 ± 5,1	1	83	48,0 ± 16,7	0	99
GENOTYPES NON-CONCORDANTS	7,6 ± 4,8	0	44	6,7 ± 2,8	0	99

Enfin, nous ne conservons, sur les variants ayant passé les filtres précédents, que ceux pour lesquels on observe au minimum un allèle alternatif sur les 73 séquences d'Alpine et Saanen restantes. Ceci permet d'écarter une grande partie des variants monomorphes dans nos races d'intérêt. Ces derniers n'ont pas d'intérêt pour l'imputation et les analyses d'association et augmentent les temps de traitement. Il reste toutefois quelques variants monomorphes dans nos races d'intérêt : i) allèles alternatifs à la séquence de référence et ii) les variants qui sont monomorphes dans chaque race prise séparément. Tous ces variants monomorphes ne seront pas pris en compte dans les calculs de qualité d'imputation et les analyses d'association qui sont réalisés séparément par race.

II.2.c. Vérification de la qualité des séquences en sortie de filtrage

Nous aboutissons donc sur le dernier filtrage à plus de 23 millions de variants dont 40 491 sont des variants présents de la puce 50k après contrôle qualité. La qualité de ce dernier filtrage a été évaluée en utilisant les génotypages 50k disponibles pour les animaux séquencés. Un taux de concordance entre 50k et séquence a été calculé pour chaque individu. Dans le cadre de cette thèse nous avons considéré que ce taux était correct quand il était situé au-dessus de 95%. En sortie, nous avons obtenu un taux de concordance moyen de 98,24%, un minimum de 94% et un maximum de 99,96%.

II.3. Imputation post-filtrage des séquences

Nous avons pu constater que suite au filtrage des données, certains individus séquencés avaient un pourcentage de génotypes manquants important (en moyenne 4.63% ± 9.59 et pouvant atteindre 66% de la séquence pour certains individus). Nous avons donc imputé les génotypes manquants à l'aide des séquences disponibles. Plusieurs logiciels d'imputation ont été envisagés. Beagle (Browning & Browning, 2011) est le logiciel d'imputation qui a été utilisé sur les séquences bovines pour corriger les génotypes en sortie de filtrage (Daetwyler et al., 2014). Nous avons donc testé Beagle après filtrage des données en imputant les génotypes manquants sur chacun des groupes de races définis dans le cadre du projet AdaptMap (Evol et al., 2018) séparément. Ce test a été effectué sur le fichier vcf contenant 594 séquences. En moyenne la concordance 50k/séquence était de 98,19% et variait entre 95,17% et 99,97% en sortie de filtrage. En sortie de Beagle, la concordance moyenne a chuté à 95,93% et variait alors entre 68,46% et 99,96%. Nous avons, par conséquent, envisagé une autre stratégie d'imputation et appliqué AlphaImpute (Antolín et al., 2017) ou FImpute (Sargolzaei et al., 2014) ou successivement les logiciels AlphaImpute et FImpute. Ces logiciels ont pour avantages principaux d'avoir des formats d'entrées similaires, d'être rapides et de pouvoir prendre en compte le pedigree pour l'imputation. Cette dernière option est intéressante car on compte parmi les Alpines et Saanen françaises séquencées 13 paires de cousins et 7 paires de parent/descendant. Il est toutefois à noter qu'il ne nous a pas été possible de lancer AlphaImpute avec le pedigree car les temps de calcul et la mémoire requise étaient trop importants. Nous avons donc imputé les séquences intra-race et obtenus les résultats présentés dans le Tableau 8.

Tableau 8: Résultats de concordance 50k/séquence après imputation des séquences filtrées par différents logiciels (vcf de 594 séquences).

	TAUX DE CONCORDANCE 50K /SEQUENCE (EN %)			***POURCENTAGE DE DONNEES MANQUANTES RESTANTES***		
	Moyenne	Min	Max	Moyenne	Min	Max
ALPHAIMPUTE (SANS PEDIGREE)	98,99	97,34	99,98	5,66	0,00	37,38
FIMPUTE	97,07	81,03	99,97	0,00	0,00	0,00
ALPHAIMPUTE (SANS PEDIGREE) + FIMPUTE	97,04	81,62	99,97	0,00	0,00	0,00

Suite à ces analyses, nous avons retenu le processus qui consiste à imputer successivement avec AlphaImpute puis FImpute. Cette imputation ne permet pas d'améliorer la concordance 50k/séquence mais elle attribue un génotype à toutes les données manquantes. De plus, par la suite, nous avions envisagé d'utiliser FImpute pour imputer les génotypes 50k vers la séquence. FImpute apporte des modifications aux génotypes des individus lorsqu'il détecte des incohérences mendéliennes. Il nous paraissait donc important de quantifier les modifications apportées aux séquences de référence par le logiciel.

II.4. Evolution du filtrage et retour sur analyses

La qualité de filtrage est mesurée par le calcul d'une concordance entre les génotypes 50k et la séquence d'un individu. Cependant, cette mesure est relativement limitée puisqu'elle permet seulement de mesurer la qualité de filtrage ponctuellement sur des marqueurs déjà sélectionnés pour leur qualité. Il nous a donc paru nécessaire d'évaluer l'imputation de la 50k vers la séquence et d'effectuer les analyses d'association sur les séquences imputées avant de revenir sur le filtrage. Ce dernier a été adapté en comparant les résultats d'imputation à la littérature disponible dans d'autres espèces d'élevage et les résultats d'analyse d'association avec les résultats précédemment obtenus en Alpine et Saanen françaises pour des caractères identiques (Martin et al., 2018; Martin et al., 2017; Oget et al., 2018; Palhière et al., 2018).

Par exemple, nous avions imposé un minimum de 3 lectures de chacun des allèles pour les génotypes hétérozygotes sans quoi le génotype était déclaré manquant. Ce filtre s'appuie sur le fait que le *calling* des hétérozygotes nécessite une grande couverture pour être correct (T. Druet et al., 2014). Nos séquences sont peu profondes et ce filtre a abouti à perdre le génotype d'un grand nombre des hétérozygotes disponibles. Ainsi, 14 séquences parmi les 81 étudiées avaient un pourcentage de génotypes inconnus supérieur à 50%. Dans ces conditions, le nombre de variants conservés était faible et bien que la qualité d'imputation soit bonne (concordances génotypes imputés/génotypes vrais entre 93,4 et 93,5%), les analyses d'association ont échoué à retrouver certains QTL connus. Par exemple, deux mutations causales ont été caractérisées sur le chromosome 14 dans le gène DGAT1 pour le TB (Martin et al., 2017). Sur ces dernières, une seule était encore présente dans le jeu de données mais n'était pas significativement associée au caractère à l'issue de l'imputation. Nous avons donc supposé que le filtrage avait écarté une grande partie des variants intéressants et que parmi les

variants restants, les génotypes hétérozygotes avaient été trop rigoureusement filtrés modifiant considérablement les fréquences des allèles.

II.5. Traitement des variants de la séquence correspondant à des marqueurs sur la puce 50k

De manière générale, le génotypage sur puce à ADN est plus exact que le séquençage en faible profondeur pour définir le génotype à une position donnée du génome car les sondes présentes sur la puce sont conçues spécifiquement avec une information a priori sur le variant. Lorsque les filtres sur les données de séquence sont stricts, la grande majorité des variants correspondant à des marqueurs de la 50k (voire la quasi-totalité) peut être perdue. Il nous a donc fallu trouver un compromis pour conserver suffisamment de ces variants présents sur la puce tout en nous assurant de ne pas propager des erreurs de séquençage dans la suite des analyses. Après vérification de la concordance 50k/séquence, nous remplaçons les génotypes des marqueurs de la puce d'un individu dans la séquence de cet individu avant imputation intra-séquencés. La qualité des imputations et la significativité des analyses d'association s'en sont trouvées améliorées. Ces dernières ont, en effet, pu atteindre le niveau de significativité obtenu avec la seule analyse des génotypes 50k alors qu'il était jusque-là inférieur dû à la moindre qualité des génotypes issus du séquençage.

II.6. Description des données en sortie de filtrage

Initialement, notre jeu de 829 séquences a permis de mettre à jour 110 193 942 variants dont près de 11,2% sont des petites insertions-délétions. Cette proportion est comparable à ce qui a pu être identifié avec un panel de 274 taureaux français séquencés (Boussaha et al., 2016). En fin de filtrage, le jeu de données n'incluait plus que 1,65 millions d'indels soit 7,1% des indels initiaux. La majorité des indels n'a pas été conservée car ces derniers ont en général une moins bonne qualité. Ils sont, en effet, plus difficiles à aligner sur un génome de référence qu'un simple SNP. En sortie de filtrage, la MAF moyenne est de 18,05% et 18,17% en Alpine et Saanen respectivement. En comparaison, dans le cadre du projet 1 000 génomes humains (Auton et al., 2015), plus de la moitié des variants présentent des fréquences très faibles (MAF < 0.5%). Cette proportion est nettement supérieure à la nôtre car les variants retenus dans le projet humain incluaient des variants avec plus de 2 allèles alternatifs ainsi que des variants structuraux. Enfin la structure d'une espèce d'élevage sous sélection est potentiellement différente de celle de l'espèce humaine qui peut être plus diverse d'un point de vue génétique. En bovins, sur un jeu de 57 615 739 variants identifiés à l'aide

de 1 147 séquences, la MAF moyenne a été estimée à 6.93%. Cette dernière est encore inférieure à celle que nous avons estimée en Alpine et Saanen française, toutefois, la moyenne a été calculée sur près de 37 races différentes (dont des races composites) incluant des races laitières comme allaitantes (Mekki Boussaha, INRAE GABI, communication personnelle).

III. Imputation des génotypes 50k vers la séquence - Article

III.1. Introduction et résumé de l'article

Le projet VarGoats est une opportunité unique d'utiliser un panel de séquences caprines sans précédent. Ces séquences incluent des boucs d'insémination largement utilisés et nous permettent d'envisager des études prospectives de l'imputation. L'imputation permettrait d'obtenir à moindre coût un grand nombre de séquences qui pourront être utilisées pour d'autres analyses. Des études similaires ont été effectuées dans d'autres filières. Ainsi, la littérature a montré qu'une imputation par pallier vers la séquence, c'est-à-dire avec une première étape d'une basse ou moyenne densité vers une haute densité puis de cette haute densité vers la séquence intégrale était plus précise et conduisait à moins d'erreurs dans les génotypes imputés (Van Binsbergen et al., 2014). En caprin, l'exercice n'est pas envisageable car la seule puce à ADN disponible est une puce de moyenne densité. Il nous faut donc imputer directement de la moyenne densité vers la séquence. Cette imputation a été étudiée sur des populations séquencées de taille similaires aux nôtres en bovins laitiers (H Li et al., 2014) et en volailles (Ye et al., 2018). Elle produit des résultats corrects puisque Li, *et al.* (2014) obtiennent des taux de concordance compris entre 0,75 et 0,85 et des corrélations comprises entre 0,63 et 0,76. Ye *et al.*, (2018) obtiennent quant à eux des taux de concordances d'environ 0,8 en fonction du chromosome étudié. Nous avons donc envisagé d'utiliser les séquences caprines disponibles pour effectuer cette imputation.

La puce caprine couvre correctement le génome et, comme souligné dans le chapitre précédent, a permis les premières identifications de QTL. Toutefois, les séquences représentent une opportunité pour affiner les signaux précédemment observés dans les caractères d'intérêt pour la filière. En effet, les données de séquence densifient la représentation du génome et contiennent potentiellement des mutations causales responsables des variations phénotypiques des caractères étudiés.

Notre objectif était d'étudier la faisabilité d'une imputation directe de la puce 50k caprine vers la séquence. Nous avons défini la stratégie optimale pour maximiser la qualité

d'imputation. Nous avons ensuite évalué la capacité des données de séquence imputées à identifier des régions d'intérêt sur le génome. Pour cela, nous avons utilisé les caractères de quantité de lait et production de semence (volume, concentration des éjaculats et nombre de spermatozoïdes).

L'imputation directe de la puce vers la séquence a été étudiée sur les Alpine et Saanen françaises en utilisant FImpute et différentes populations de référence : individus séquencés de la race, toutes les séquences *Capra hircus* disponibles, chèvres laitières européennes, chèvres de France métropolitaine. L'imputation de meilleure qualité a été obtenue en utilisant le panel de séquences de France métropolitaine avec des taux de concordance moyens de 0,86 et 0,75 et des corrélations de 0,26 et 0,24 respectivement en Alpine et Saanen. Toutefois les différences entre les scenarii sont minimes. Nous avons tout de même noté un effet du scenario d'imputation utilisé sur la détection de régions d'intérêt avec les analyses d'association. De nouveaux signaux ont été détectés pour la production laitière sur les chromosomes 2 et 5 en Alpine et Saanen en utilisant les séquences imputées à l'aide du panel de séquence françaises. Deux signaux ont également été identifiés pour le volume de semence et la quantité de lait dans une région du chromosome 19 comprise entre 23 et 28 Mb. Un CLIP test (Close LInkage versus Pleiotropy) a été effectué sur les variants de la région et a réfuté l'hypothèse de pléiotropie.

Etant données les faibles différences d'un scenario à l'autre, l'utilisation d'une imputation intra-race s'avère plus efficace car elle est moins exigeante en temps de calcul et préparation des données. De plus, l'imputation utilisant les séquences françaises a fortement augmenté le bruit sur les analyses d'association en Saanen et légèrement réduit la significativité des signaux. Cependant le faible nombre de séquences disponibles intra-race est fortement limitant et ne permet probablement pas la détection de plus petits signaux. Une augmentation du nombre de séquences disponibles dans ces races est donc obligatoire si une recherche plus approfondie de signaux additionnels est envisagée. En parallèle, l'addition de variants significatifs à une puce de génotypage permettrait probablement de confirmer les signaux identifiés suite à l'imputation vers la séquence.

III.2. Analyses d'association pour les caractères de production de semence et de quantité de lait utilisant différentes stratégies d'imputation vers la séquence en caprins laitiers français - Article

Talouarn et al. BMC Genetics (2020) 21:19
https://doi.org/10.1186/s12863-020-0826-9

BMC Genetics

RESEARCH ARTICLE Open Access

Genome wide association analysis on semen volume and milk yield using different strategies of imputation to whole genome sequence in French dairy goats

Estelle Talouarn[1*], Philippe Bardou[1,2], Isabelle Palhière[1], Claire Oget[1], Virginie Clément[3], The VarGoats Consortium, Gwenola Tosser-Klopp[1], Rachel Rupp[1] and Christèle Robert-Granié[1]

Abstract

Background: Goats were domesticated 10,500 years ago to supply humans with useful resources. Since then, specialized breeds that are adapted to their local environment have been developed and display specific genetic profiles. The VarGoats project is a 1000 genomes resequencing program designed to cover the genetic diversity of the *Capra* genus. In this study, our main objective was to assess the use of sequence data to detect genomic regions associated with traits of interest in French Alpine and Saanen breeds.

Results: Direct imputation from the GoatSNP50 BeadChip genotypes to sequence level was investigated in these breeds using FImpute and different reference panels: within-breed, all *Capra hircus* sequenced individuals, European goats and French mainland goats. The best results were obtained with the French goat panel with allele and genotype concordance rates reaching 0.86 and 0.75 in the Alpine and 0.86 and 0.73 in the Saanen breed respectively. Mean correlations tended to be low in both breeds due to the high proportion of variants with low frequencies.

For association analysis, imputation was performed using FImpute for 1129 French Alpine and Saanen males using within-breed and French panels on 23,338,436 filtered variants. The association results of both imputation scenarios were then compared. In Saanen goats, a large region on chromosome 19 was significantly linked to semen volume and milk yield in both scenarios. Significant variants for milk yield were annotated for 91 genes on chromosome 19 in Saanen goats. For semen volume, the annotated genes include YBOX2 which is related to azoospermia or oligospermia in other species. New signals for milk yield were detected on chromosome 2 in Alpine goats and on chromosome 5 in Saanen goats when using a multi-breed panel.

Conclusion: Even with very small reference populations, an acceptable imputation quality can be achieved in French dairy goats. GWAS on imputed sequences confirmed the existence of QTLs and identified new regions of interest in dairy goats. Adding identified candidates to a genotyping array and sequencing more individuals might corroborate the involvement of identified regions while removing potential imputation errors.

Keywords: Sequence data, Imputation, Semen, Milk yield, GWAS analysis, French Alpine and Saanen, Goats

* Correspondence: estelle.talouarn@inrae.fr
[1]GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan, France
Full list of author information is available at the end of the article

Background

The recent decrease in sequencing costs has made it possible to sequence large numbers of individuals in key livestock species. The VarGoats resequencing program is the logical next step following the ADAPTmap initiative on 50 k genotyping data [1]. The program plans to sequence over 1000 animals of the *Capra* genus from 65 breeds including 44 French Alpine and 37 French Saanen animals. The sequenced individuals include widely used AI bucks and and enabled to perform a preliminary analysis of imputation for subsequent association analyses. They cover an appreciable part of the effective population sizes of both breeds, estimated to be 115 and 98 in Alpine and Saanen respectively (Carillier, 2015, INRAE, personal communication).

Imputation is a more cost-effective method for obtaining a large amount of sequence data for subsequent analysis. A high-density genotyped reference panel is used to predict high-density genotypes in a low-density genotyped population. When possible, imputation to a whole-genome sequence is performed in a stepwise manner starting with the lowest density panel, before moving on to a medium density chip, then a high-density chip and finally imputation to sequence level. In dairy cattle and sheep, this method has proved more efficient than direct imputation from lowest density to sequence [2, 3]. In goats, the only genotyping tool available is a 50 k-chip (Illumina GoatSNP50 BeadChip) [4]. This means that imputation must be carried out directly from 50 k to sequence level.

Genome-wide association studies (GWAS) are commonly used to unravel the genetic architecture of complex traits. The GoatSNP50 BeadChip has led to the detection of a few Quantitative Trait Loci (QTL) regions for milk and type traits in French Alpine and Saanen breeds [5–8]. One causal mutation has been identified [5] and a large zone on chromosome 19 needs to be refined in Saanen goats given the width of the confidence interval, the multiplicity of traits associated with the same region of the chromosome and as no straightforward functional candidate gene was identified [6]. The use of sequence data, rather than chip data, for fine QTL mapping has proved more accurate in various species, such as cattle [9, 10] and poultry [11]. Indeed, chip data consist of only a few variants selected based on their quality (length of the contig, proximity to other SNPs, exclusion of tri-allelic and A/T or C/G SNPs, estimated quality of the probe etc. ...), spacing and MAF, therefore variants with low MAF are under-represented. However, rare variants could actually have a significant impact on the phenotypes studied as they might have appeared only recently in the target populations. Sequence data include various MAF profiles and should contain the causal mutations that affect the traits of interest. It is therefore preferable to perform association analyses on the whole-genome sequence (WGS) rather than on chip data that mainly rely on linkage disequilibrium with the nearby causal mutation. Besides, chip data may only contain SNPs (single nucleotide polymorphisms) whereas sequence data include both SNPs and small indels (insertion/deletion).

Few association studies have been conducted on semen production traits in goats. For example, Nickbin et al. [12] investigated the HSP70 gene in Boer goats and Mohammed et al. [13] calculated genetic parameters for the Damascus breed. However, the association of semen production traits to regions of the genome has yet to be investigated in Alpine and Saanen breeds despite their economic importance in the French dairy industry. In France, where around 70,000 artificial inseminations are performed every year with Alpine and Saanen bucks, semen production traits are of major interest. Bucks culled for semen defects represent a burden for the French breeding organization, CapGenes. Indeed, nearly 46% of the 120 to 130 young bucks that enter the progeny testing process are discarded due to semen quality issues.

This study is the first to investigate imputation in dairy goats. Our main objectives were to evaluate the quality of imputation, define the best imputation scenario and finally assess the usefulness of imputed sequence data to identify genome regions associated with semen production and milk yield traits in French Alpine and Saanen breeds.

Methods

Data available

No animal experiments were necessary for this study, therefore no ethics committee approval was required. Sequence data were obtained from the VarGoats project using Illumina HiSeq or Illumina NovaSeq technologies, the first step towards a 1000 goat genome project (http://www.goatgenome.org/vargoats.html). The current data bank comprises 808 individuals from *Capra hircus* of various breeds and geographical origins, as well as 21 wild goat individuals. Forty-four French Alpine and thirty-seven French Saanen individuals were sequenced at Genoscope (Evry, France) with an average coverage of 12X. The individuals sequenced were selected to best represent the genetic structure of the current French population: AI bucks from the largest families, maximized haplotype coverage from the population by picking unrelated individuals following the approach described by Druet et al. [14]. However for some research purposes (milk flow trait, specific casein profiles), a few closely related individuals were sequenced and added to the overall dataset. Thus, sequence data include 13 pairs of cousins ($G_{jk} > 0.12$) and 7 parent/descendant

pairs ($G_{jk} > 0.40$). All selected animals except 3 had previously been genotyped using the Illumina GoatSNP50 BeadChip.

A total of 2455 French Alpine individuals (994 males, 1461 females) and 1570 French Saanen individuals (757 males, 813 females) genotyped using the Illumina GoatSNP50 BeadChip were available for imputation. Pedigree information was available and used as the genotyped individuals were closely related to the reference panel of sequences. Data were cleaned using an in-house pipeline as described in Martin et al. (2018). In brief, all individuals with a call rate below 95% or showing pedigree inconsistency (i.e. having more than 10% parent/offspring conflicting SNPs) were discarded. SNP quality control was based on the following inclusion criteria: call rate above 99%, MAF above 1% and Hardy-Weinberg *P*-value above 10^{-6}. After editing, a total of 47,147 synthesized SNPs (out of a total of 53,347) remained on goat autosomes CHI 1 to CHI 29 and were used for subsequent analyses. Marker orders and positions were based on the ARS1 caprine Assembly [15] . The GoatSNP50 BeadChip SNP positions were updated on ARS1 genome assembly as described on the VarGoats website (http://www.goatgenome.org/projects.html) and made publically available by the International Goat Genome Consortium.

Sequence data quality check and imputation

The sequenced reads were aligned to the goat reference genome assembly ARS1 (https://www.ncbi.nlm.nih.gov/assembly/GCF_001704415.1/) using the Burrows–Wheeler Alignment tool (BWA-MEM version 0.7.15) with default parameters [16].

According to GATK best practices, BAM files were preprocessed: duplicates were removed, indels realigned and base quality score recalibrated with Picard tools version 2.1.1 and Genome Analysis Toolkit (GATK) version 3.7–0 [17]. Variant calling was performed for all GVCF files using GATK HaplotypeCaller and variants were annotated using SnpEff (version 4.3 t) [18] and the NCBI *Capra hircus* annotation release 102 (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/704/415/GCF_001704415.1_ARS1/).

Variant calling on the 829 individuals led to the identification of 110,193,942 variants on the 29 autosomal chromosomes: 97,889,899 SNPs and 12,304,043 indels. Among the 829 sequenced individuals, 16 had a mean coverage below 5 (4 French Alpine and 4 French Saanen) and were removed from the data set for subsequent analyses. The sequence dataset therefore consisted of 815 sequenced individuals including 40 French Alpine goats (36 males and 4 females) and 33 French Saanen goats (31 males and 2 females). Quality checks were applied to the sequence variants using the indicators listed in Fig. 1. The thresholds for individual genotype quality (GQ) and individual depth (DP) were set to 8 and 7 respectively by comparing the genotypes of the GoatSNP50 BeadChip SNPs with the sequence variants. The mean GQ and mean DP were 6.7 (± 2.8) and 7.6 (± 4.8) respectively for mismatching SNPs compared with 48.0 (± 16.7) and 11.7 (± 5.1) respectively for the matching genotypes.

After the quality check, only variants with at least one observation of the alternative allele (ALT) in French Alpine or Saanen animals were retained in order to reduce computation time in subsequent analyses and only keep variants of interest in our breeds. Thus, 23,338,436 variants, including 40,491 GoatSNP50 BeadChip SNPs, were kept for imputation. Concordance with the 50 k genotypes was checked. After variant filtering, the individual mean concordance rate was 98.24% (± 1.12) and ranged from 94.00 to 99.96%.

Imputation of missing genotypes in the sequence reference panel

It should be emphasized that missing genotypes represented on average 4.63% of all the sequence variants for an individual of French Alpine and Saanen breeds. This percentage could attain 66% if sequencing was of low coverage. A within-breed imputation was therefore applied to fill in the gaps. Using a combination of AlphaImpute (v 1.9) [19] and FImpute (v 3.0) [20] gave higher concordance rates than using solely one software while minimizing computation time. We hence imputed filtered sequences using AlphaImpute and FImpute consecutively for French Alpine and Saanen breeds separately. The mean concordance rate between 50 k genotypes and sequence data was 98.62% (± 1.19) after imputation and no missing genotypes remained. For subsequent analyses, as chip genotypes are more reliable than low-depth sequencing, and to avoid spreading

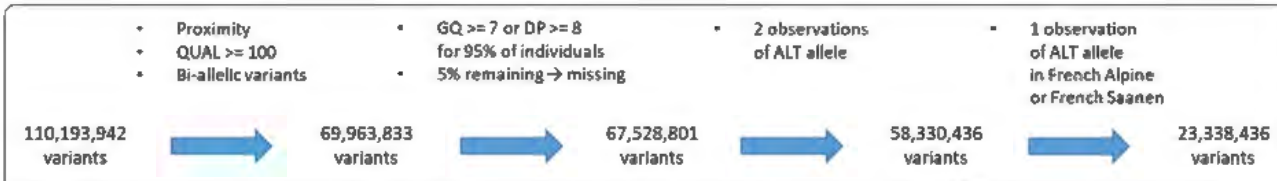


Fig. 1 Filtering process of sequence data from 829 individuals (VarGoats 1000 genomes project) including 33 French Saanen and 40 French Alpine goats. QUAL: quality estimated by GATK, GQ: individual genotype quality, DP: individual depth, ALT: alternative allele

genotyping errors down the pedigree, the 50 k markers in the sequencing data were systematically replaced by information from 50 k genotypes, when available.

Animal phenotypes

Male traits

Three semen production traits were recorded on artificial insemination (AI) bucks (Table 1): semen volume in mL (SV), semen concentration in billions of spermatozoa per mL (SC) and number of spermatozoa in billions of spermatozoa (SN). Semen production and quality were analyzed in 305,840 ejaculates from 2865 AI bucks from the CapGenes breeding organization (Mignaloux-Beauvoir, France). Mean yield deviations (YD) per buck were computed from repeated performances (1 to 447 repetitions per buck) then corrected for environmental effects: age, month and year of semen collection, and time between two consecutive samples.

Female traits

Milk yield (MY) in kg was also measured as detailed by Martin et al. [5, 6] and analyzed. Daughter Yield Deviations (DYD) were computed for males with at least 10 daughters with records (Table 1). DYDs were the average daughters' performances corrected for environmental effects and merit of the dam.

Imputation scenarios of 50 k genotypes to sequence level and quality assessment

Imputation of 50 k genotypes to sequence level was performed using FImpute software (v 3.0) which takes pedigree information into account [20]. The accuracy and efficiency of FImpute, compared with various other imputation tools, has been confirmed [21–23]. Imputation quality was checked before imputing the available 50 k genotypes to sequence level. A leave-one-out scenario was applied to 4 sequenced daughters of 2 different sequenced sires (2 Alpine and 2 Saanen) to maximize the kinship with the reference population. One of the daughters was in turn masked down to a 50 k-equivalent and then imputed. The allele and genotype concordance rates (CR) and Pearson correlations (R) of the true and imputed sequences were then calculated per variant and per MAF profile.

Various reference populations were tested based on their proximity to French Alpine and Saanen goats. To accurately build the different reference panels, a Principal Component Analysis (PCA) was performed on chromosome 1 filtered data using PLINK software [24]. Groups were then formed based on the origin of an individual and on the PCA results (Fig. 2). The number of individuals per group is given in Table 2. All sequenced wild individuals were removed from reference populations as they are genetically different from the *Capra hircus* species. We also tried to impute sequences without pedigree while using all sequenced French goats available (excluding Angora and Creole breeds).

Table 1 Available phenotypes for association analysis. Semen production traits included spermatozoa number, semen concentration and semen volume

	Alpine	Saanen
Milk yield trait (DYD of AI bucks)	631	483
Semen production traits (YD of AI bucks)	668	515

Association analysis

The imputed sequences were subjected to single-trait association analysis for milk and semen production traits using mixed linear models with the *mlma* option of GCTA software [25] and the following model:

$$y = 1\mu + xb + u + e$$

where **y** represents pre-adjusted phenotypes of the trait; μ is the overall mean; b is the additive fixed effect of the variant tested; $\boldsymbol{x}$ is the vector of imputed genotypes coded in 0, 1, 2 (copy number of the alternative allele); $\boldsymbol{u}$ is the vector of random additive polygenic effects, $\boldsymbol{u} \sim N(\mathbf{0}, G\sigma^2)$ with **G** the genomic relationship matrix; $\boldsymbol{e}$ is the vector of random residual effects normally distributed. The genomic relationship **G** matrix was calculated on 50 k genotypes using PLINK [24].

The four traits were subjected to within-breed association analysis (Table 1). Variants with a within-breed MAF lower than 1% were excluded, leaving 11,933,965 and 12,449,740 variants in Alpine and Saanen goats, respectively, when sequences were imputed within-breed, and 14,695,413 and 15,404,361 variants in Alpine and Saanen goats, respectively, when imputation was performed using the French multi-breed panel. A Bonferroni correction was applied to the significance thresholds to account for multiple testing. The average chromosomal significance level was calculated as follows: $-log_{10}(0.05/(number\ of\ variants/29))$.

The results of the sequence data association analysis were then compared with 50 k-genotypes results, by performing a GWAS on the 40,491 SNPs found both in the filtered sequence data and the cleaned GoatSNP50 BeadChip SNPs.

Annotations were extracted from VCF files for variants with a $-\log_{10}(p\text{-value})$ above the chromosomal threshold. The RumimiR database (http://rumimir.sigenae.org/) [26] was also checked for miRNAs located close to a significant variant.

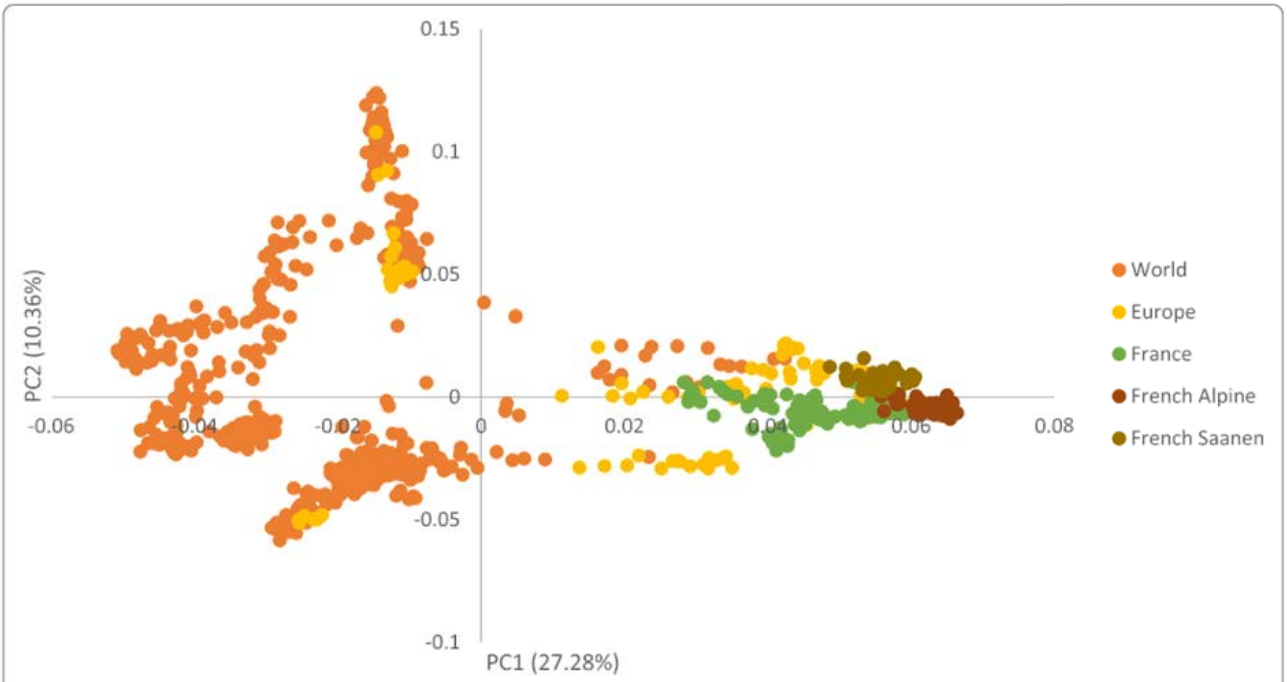


Fig. 2 Principal Component Analysis performed on 1,546,528 filtered variants on chromosome 1 using Plink software. The 'World' group comprises all sequences available except for wild animals. The 'Europe' group comprises all sequences from dairy goats from European countries (therefore excluding Boer, Angora and Creole breeds). The 'France' group comprises all sequences of goats from France (excluding Angora and Creole breeds)

Results

Imputation accuracy

Allele and genotype concordance rates (CR) and correlations (R) between true and imputed sequences were computed for all chromosomes separately. The mean allele CR, genotype CR and R were calculated per variant and per group of variants with the same MAF. Results per MAF are shown in Fig. 3 for within-breed imputation and imputation with all sequenced French goats. As shown in Fig. 3, imputation using a French multi-breed panel performs slightly better than with a breed-specific reference panel for a specific MAF, regardless of the breed. However, when considering the overall results (Table 3), the difference between the two imputation scenarios is less obvious than when comparing a MAF profile. The high proportion of low MAF in our data tended to flatten the differences as imputation quality measurements are similar in both scenarios for low MAF (from 34 to 38% depending on the breed, Fig. 4). In Saanen, the European multi-breed panel performed better. However, the difference with the French multi-breed imputation is minimal and the computation time increased with the number of sequenced individuals. We therefore chose the less time-consuming multi-breed scenario for further comparisons with within-breed imputation.

Table 2 Composition of the different reference populations used for imputation. Details of breed composition available on: http://www.goatgenome.org/vargoats.html

	Number of individuals	
	Alpine	Saanen
Within-breed	39	32
World	793	
Europe	243	
France	169	

GWAS analysis

Milk yield

In the Alpine breed, when sequence data was imputed using solely data from French Alpine, only 3 variants out of 11,933,965 reached the chromosomal significance level ($p-value \leq 1.22 * 10^{-7}$) for milk yield (Table 4) and no clear signal was detected (Fig. 5). When imputing all 50 k genotypes to sequence level using all sequenced French goats (multi-breed), 9 sequence variants out of 14,695,413 reached the chromosomal significance level ($p-value \leq 9.87 * 10^{-8}$) (Table 4) and a clear signal appeared on chromosome 2 between 28.87 and 28.89 Mb (Fig. 5).

In the Saanen breed, when imputing available 50 k genotypes using only data from sequenced French Saanen, 313 variants out of 12,449,740 reached the chromosome significance level ($p-value \leq 1.17 * 10^{-7}$) for milk yield (Table 4), all of which were situated on chromosome 19 between 23.55 and 27.68 Mb. When using a French

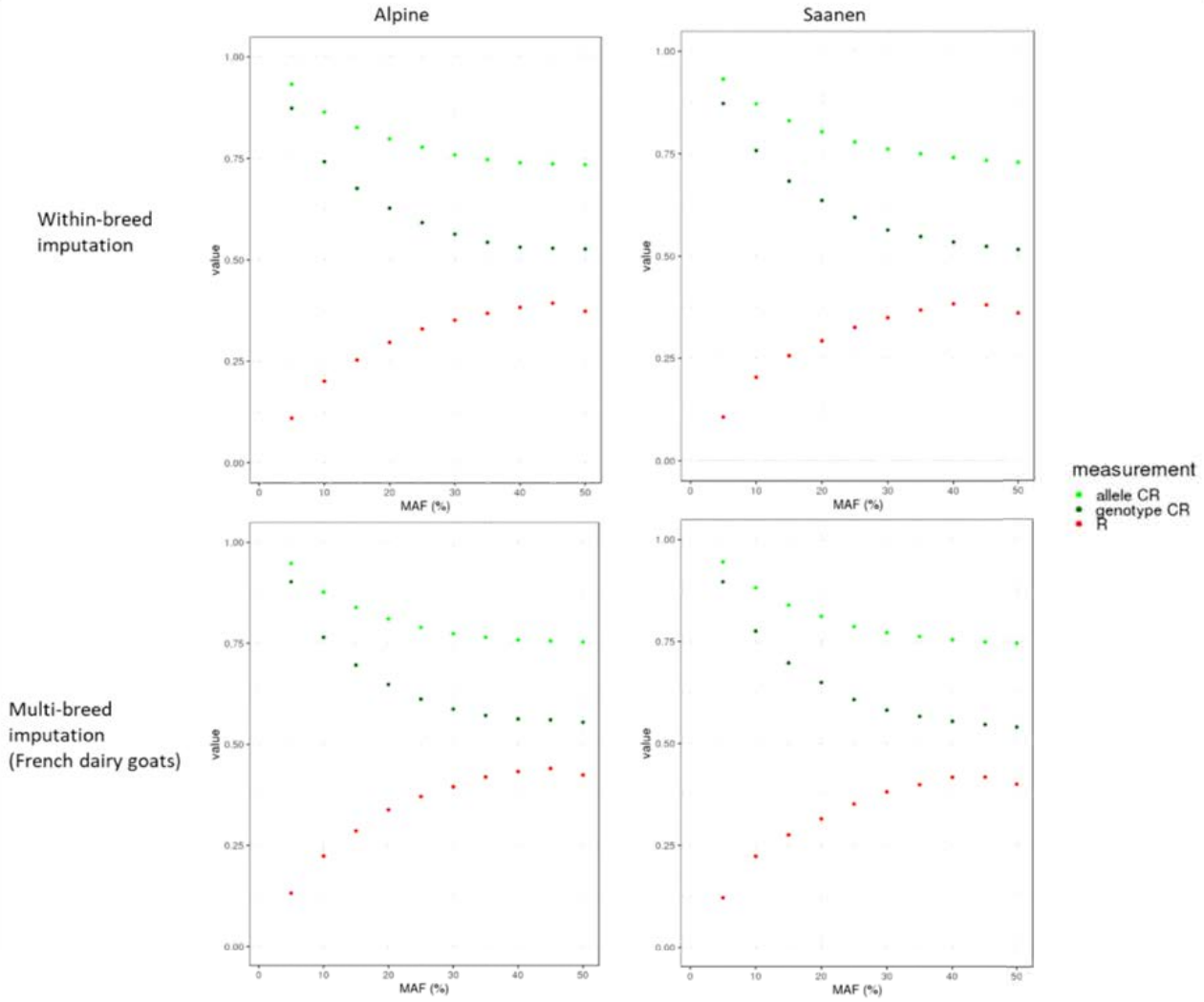


Fig. 3 Precision of imputation to sequence level per MAF for within-breed and multi-breed analysis. The precision of imputation was defined as genotype and allele concordance rates (CR) as well as correlation

multi-breed imputation reference panel, 448 variants out of 15,404,361 reached the chromosomal significance level ($p - value \leq 9.41 * 10^{-8}$) (Table 4) including 441 on chromosome 19 between 24.70 and 28.15 Mb and 7 on chromosome 5 between 44.80 and 44.81 Mb (Fig. 5).

Semen production

In the Alpine breed, no clear signal was observed for semen production traits with sequences imputed either from sequenced Alpine individuals or the French goat panel (Fig. 6).

Table 3 Correlation (R) and concordance rates (CR) between imputed and true genotypes for Alpine and Saanen breeds using different reference populations

Reference population	Pedigree	Alpine			Saanen		
		R	genotype CR	allele CR	R	genotype CR	allele CR
Within-breed	Yes	0.264	0.755	0.867	0.239	0.741	0.859
World	Yes	0.232	0.723	0.850	0.226	0.714	0.845
Europe	Yes	0.264	0.747	0.863	0.251	0.733	0.856
France	Yes	0.265	0.749	0.864	0.248	0.734	0.856
	No	0.211	0.734	0.856	0.198	0.719	0.847

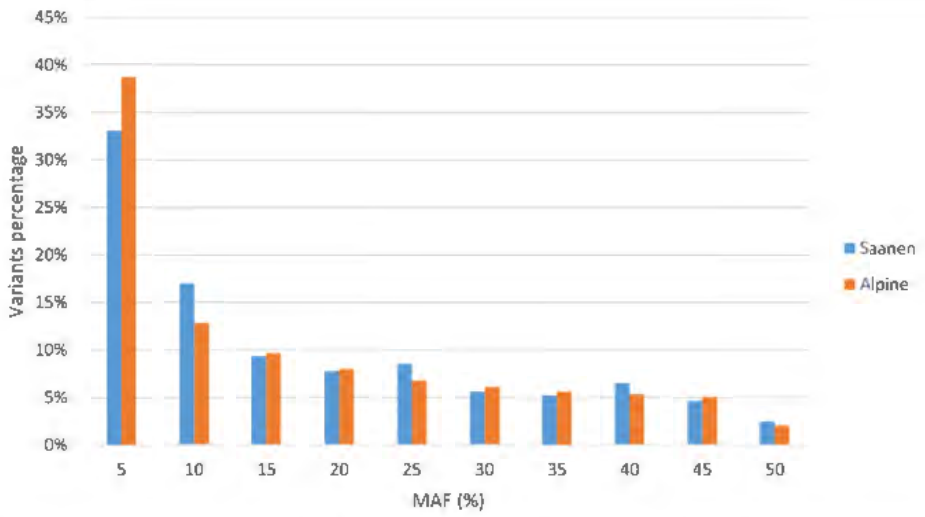


Fig. 4 MAF distribution after imputation of the 23,338,436 sequence variants retained after filtering in Alpine and Saanen goats

In the Saanen breed, a wide significant signal was found on chromosome 19 using within-breed imputation, spanning a region from 24.5 to 27 Mb. The signal was most significant for semen volume for which 209 variants reached the chromosome significance level (Table 4). However, 206 other variants were found to show significant association with this trait on the rest of the genome. When imputing the available 50 k genotypes in Saanen individuals using French goat sequences, 981 variants reached the significance level for semen volume genome-wide but only 23.8% were found on chromosome 19 (Fig. 6). A small signal was also observed for SN on chromosome 19 when using a multi-breed panel, however out the 51 genome-wide significant variants (Table 4) only 4 were located on chromosome 19.

Comparison with 50 k genotypes

The improvements provided by the imputed sequences can be easily assessed as 50 k markers genotypes (40,491 SNPs found in both the 50 k and sequence data) were directly replaced in sequence data using information of 50 k genotypes. They, therefore, underwent the same analysis using the same model, method and phenotypes. The significance levels tended to be higher with sequence data for all traits in the QTL regions (Fig. 7). Indeed, in the Alpine breed, sequence variants were systematically more significant than 50 k SNPs. In the Saanen breed, sequences variants were more significant than 50 k genotypes in every situation except for the semen volume trait when using a multi-breed reference panel for imputation. The sequence data also gave more refined peaks and a higher number of significant variants (Table 4).

Close linkage versus Pleiotropism test

As strong signals were detected for both semen volume and milk yield in the same region of chromosome 19. A Close Linkage versus Pleiotropism (CLIP) test as developed by David et al. [27] was implemented. We applied the CLIP test to sequence data imputed within-breed to determine whether the region is truly pleiotropic or if QTLs are physically close. We extracted 32,029 imputed sequence variants of chromosome 19 between 23 and 28 Mb to perform the analysis. The same analysis was performed on 50 k genotypes of the French Saanen breed and the test was detailed by Martin et al. [6]. The test

Table 4 Number of significant variants identified at the chromosome significance level in a population of 483 Saanen and 629 Alpine individuals for both imputation scenarios

	Alpine				Saanen			
Imputation	Within-breed[1]		French goats[2]		Within-breed[3]		French goats[4]	
	sequence	50 k	sequence	50 k	sequence	50 k	sequence	50 k
milk yield	3	0	9	0	313	14	448	12
number of spermatozoa	0	0	1	0	5	1	51	1
semen concentration	2	0	2	0	2	0	8	0
semen volume	2	2	2	0	415	11	981	9

Bonferroni thresholds: [1] $p-value \leq 1.22 * 10^{-7}$ [2] $p-value \leq 9.87 * 10^{-8}$ [3] $p-value \leq 1.17 * 10^{-7}$ [4] $p-value \leq 9.41 * 10^{-8}$

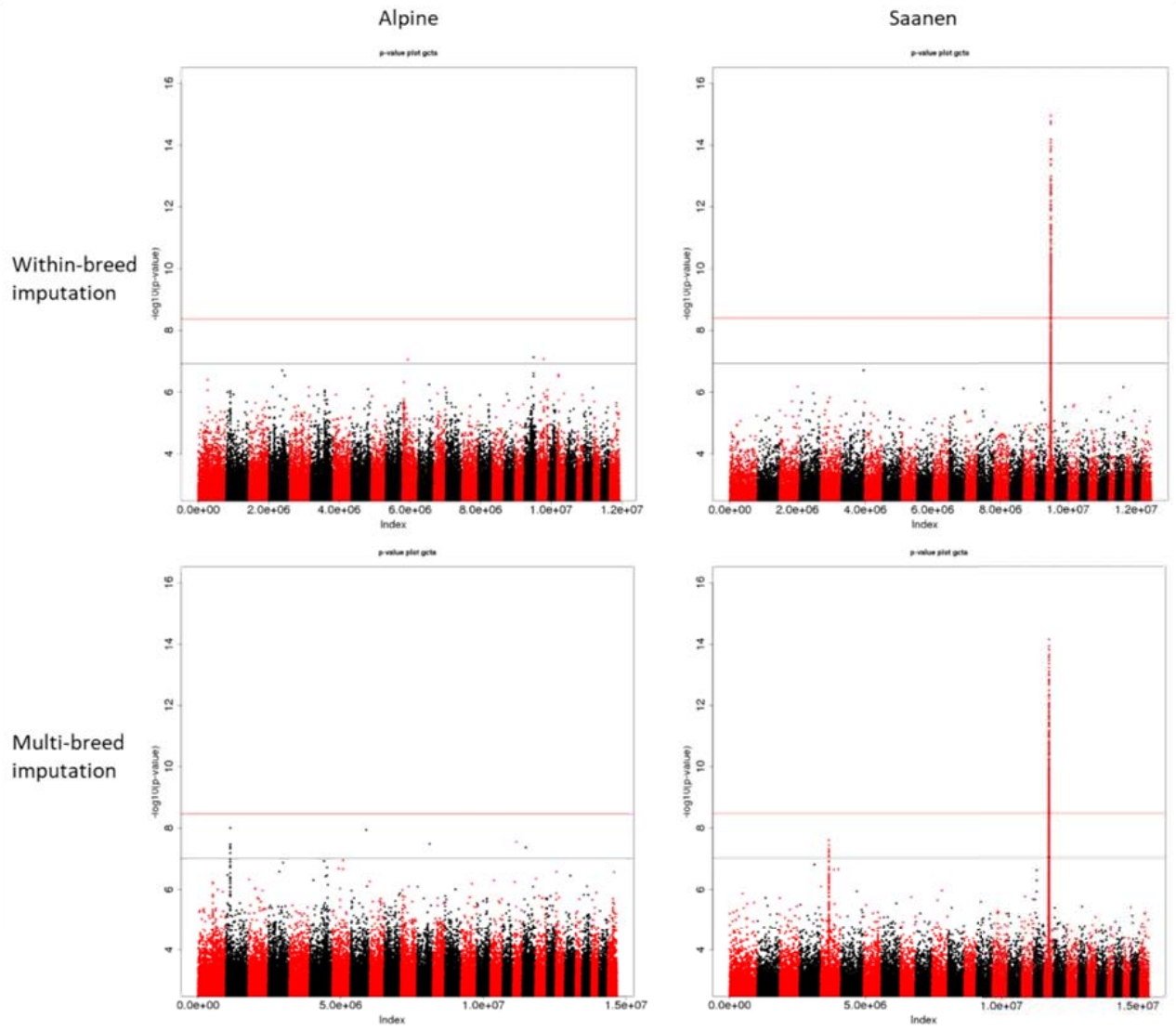


Fig. 5 Manhattan plots of genome-wide association studies (GWAS) for milk yield performed on 631 Alpine and 490 Saanen AI bucks for within-breed analysis after within-breed and multi-breed imputation. Chromosome significance threshold in black; genome-wide significance threshold in red

compares two traits X_1 and X_2 and rejects the pleiotropy if the squared correlation between a combination of effects at the variant level ($\rho^2_{X_1X_2}$), is below the minimal value it can take under the pleiotropy assumption multiplied by a factor K_α. K_α is the αth percentile of the distribution of the ration of the square of the observed correlation to its minimal value under the pleiotropy assumption.

$$\rho^2_{X_1X_2} < K_\alpha \left[\sqrt{\left(1 - \frac{1}{2N}\frac{\sigma^2_{y_1}}{\sigma^2_{X_2}}\right)} - \frac{1}{2N}\,\frac{\sigma^2_{y_1}\sigma^2_{y_2}}{\sigma^2_{X_1}\sigma^2_{X_2}} \right]$$

Where N is the number of animals included in the analysis, $\sigma^2_{X_1}$ and $\sigma^2_{X_2}$ are the observed variance of the traits and $\sigma^2_{y_1}$ and $\sigma^2_{y_2}$ are the variance of raw data.

Discussion

Imputation quality

In our study, we obtained similar imputation results with both a within-breed reference panel and a multi-breed sequence reference panel, provided that the breeds included in the reference population are quite similar to the imputed breeds (Fig. 2). We assume that using a wider reference panel covers best the genetic variability of the breed than the very little number of sequenced individuals of the breed in the VarGoats project.

Removing pedigree information before imputation strongly deteriorated the correlations (R), which decreased by 5.1 to 5.4% depending on the breed. CRs were less impacted, showing a decrease of 0.9 to 1.5% depending on the breed. In our conditions, a complete

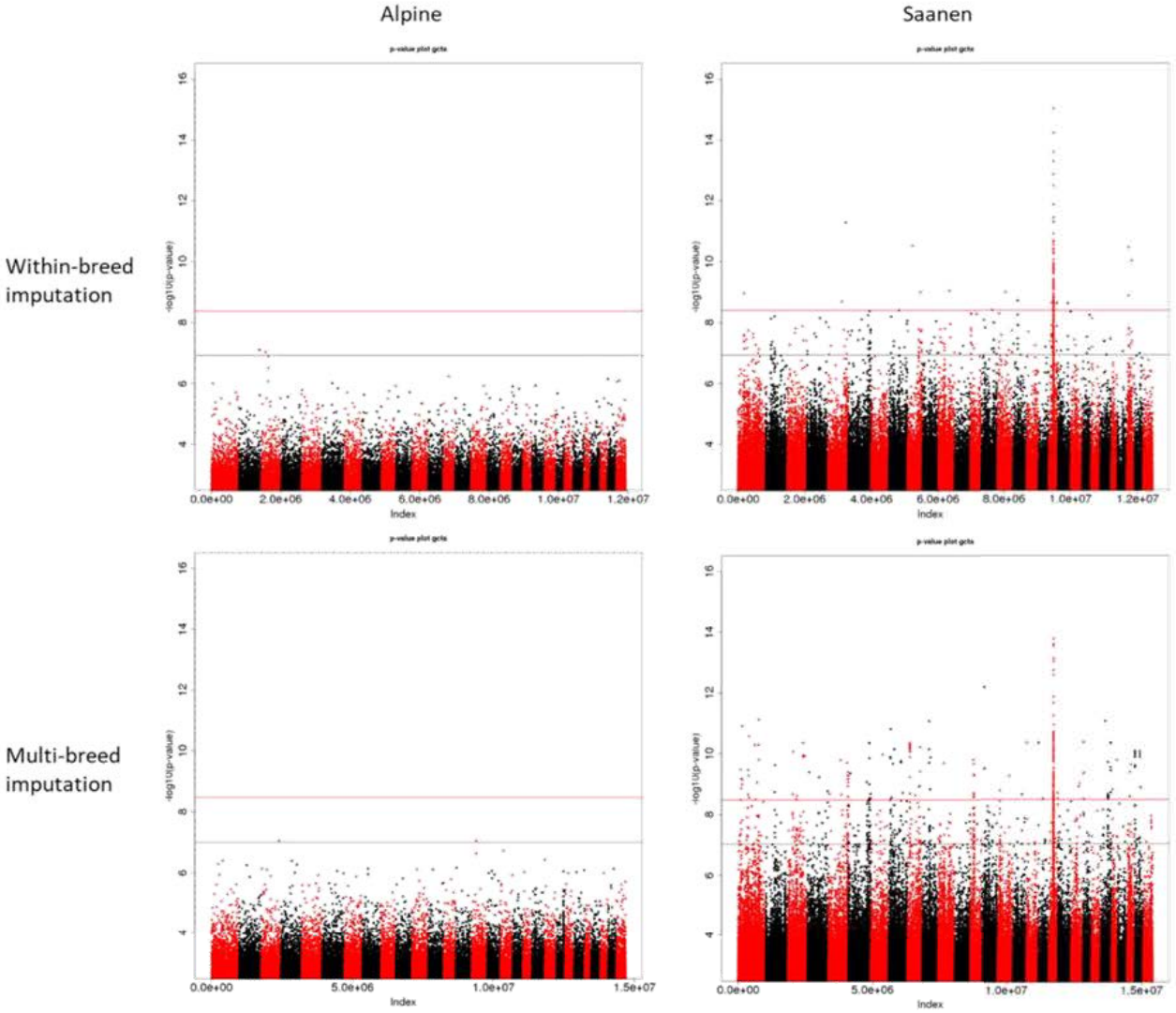


Fig. 6 Manhattan plots of genome-wide association studies (GWAS) for semen volume performed on 631 Alpine and 490 Saanen AI bucks for within-breed analysis after within-breed and multi-breed imputation. Chromosome significance threshold in black; genome-wide significance threshold in red

pedigree therefore seems useful to improve the accuracy of the imputed genotypes.

Even though the CRs obtained in this study were similar to those involving equivalent reference population sizes in other species, the correlations were significantly lower than in dairy cattle [28] or poultry [29]. Indeed, CRs ranged from 0.75 to 0.85 in Li et al. [28] and the genotype CR was around 0.8, depending on the chromosome, in Ye et al. [29]. However, the squared correlations for cattle breeds with similar population sizes ranged from 0.63 to 0.76 in Li et al. [28]. Binsbergen et al. [30] obtained results similar to ours with a reference population size of nearly 46 Holstein individuals. They performed direct imputation from 50 k genotypes to sequence level and obtained a mean correlation of 0.37 between true and imputed sequences.

One reason explaining our low correlations could be a negative effect of the large number of variants with a low MAF in our dataset (Fig. 4), for which the correlations decrease rapidly at lower MAF values (Fig. 3). Also, a slight drop in R was observed for variants with a MAF of 0.50 in Saanen goats (Fig. 3). This could be linked to the small number of variants with a high MAF (Fig. 4), and thus implies that an imputation error would have a major impact on the final correlation. Nonetheless, the correlations that we obtained even for high MAF were low in comparison with other studies in livestock [3, 29, 30] or humans [31] where the correlation between true and imputed genotypes could reach 0.8. As our imputation study involved very few sequenced individuals (33 and 40), a single imputation error would drastically reduce this correlation.

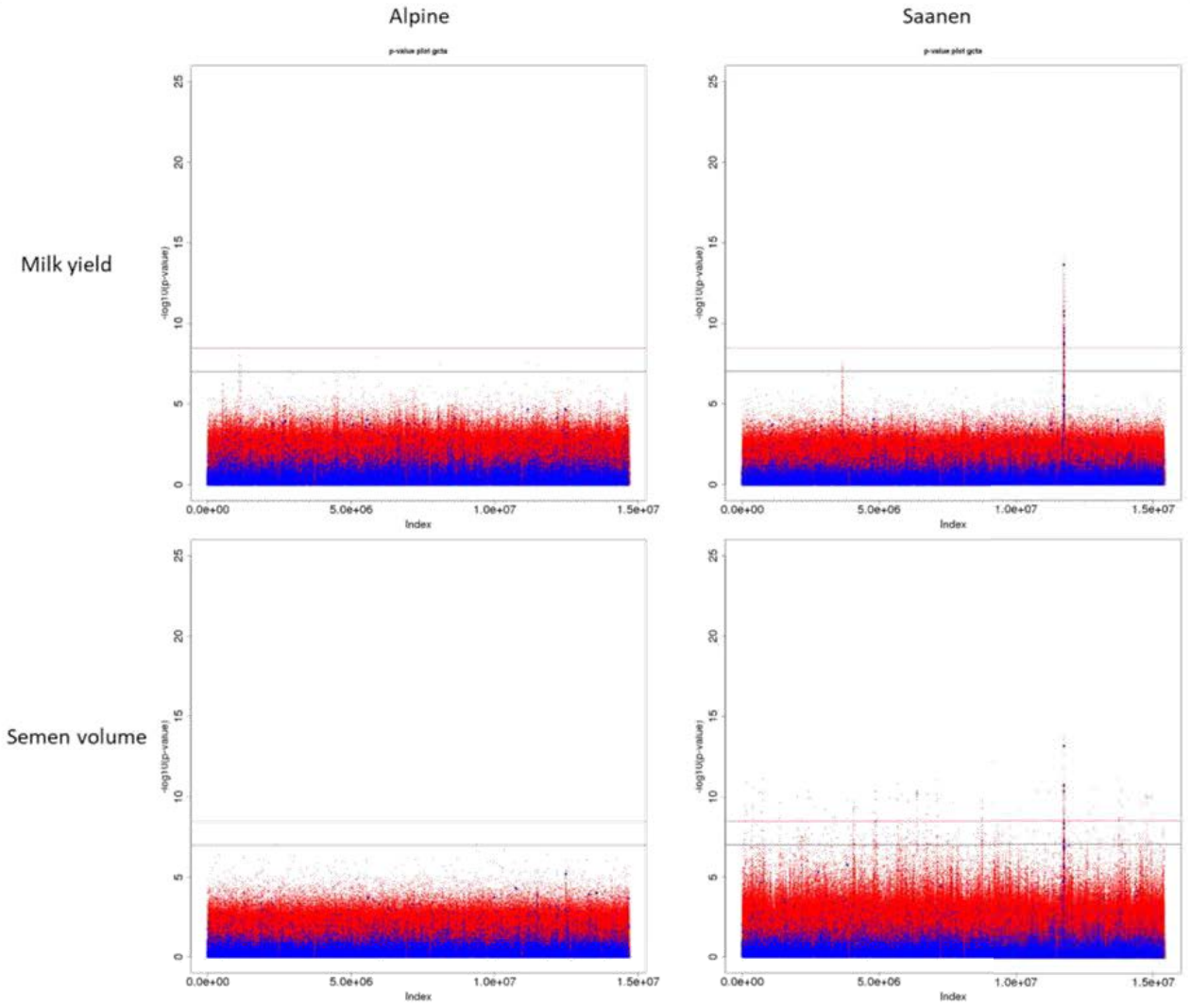


Fig. 7 Manhattan plots for milk yield and semen production after multi-breed imputation of sequences using a French dairy goats panel. Chromosome significance threshold in black; genome-wide significance threshold in red. GoatSNP50 BeadChip variants in blue; sequence variants in red

Another possible explanation is the considerable genetic diversity of the *Capra hircus* species [1, 32]. According to the French Varume project, the number of ancestors contributing to 50% of the gene pool is higher in French Alpine and Saanen (16 and 15 respectively) than in French dairy cow breeds: 7 in Montbéliarde and Holstein, 8 in Normande (Danchin-Burge, Institut de l'Elevage, personal communication). Besides, as shown by Carillier et al. [32], the LD is lower in French goats than in dairy cattle. A low LD might make phasing more difficult as the distance to a 50 k marker and therefore the number of potential recombinations increase, leading to imputation errors.

Moreover, the reference populations used for imputation were small and some individuals had initial low-depth sequences (coverage <10X). Parts of their sequence genotypes remain uncertain.

Nonetheless, most of the QTL regions previously identified with GoatSNP50 BeadChip data were detected on the sequence data with refined signals and increased significance, which suggests that the imputed sequence could be suitable for association analysis.

Nevertheless, the significance of the detected regions or the identification of new regions could be further increased by improving imputation quality. As only a limited number of animals in each breed were sequenced, some genotyping errors might be erased by increasing this number. According to Druet et al. [14], at a given sequencing effort, it is preferable to sequence more individuals at a lower coverage to detect rare variants and to call genotypes accurately than to sequence deeply few individuals.

Association analysis

The p-values for sequence data were slightly higher than for 50 k data and clear distinct signals were identified when imputed sequence data were used for association analyses (Fig. 7). Sanchez et al. [10], conducted similar studies on the Montbéliarde dairy cattle breed and reported a major increase in the significance of the detected signals when using sequence data, rather than HD or 50 k genotypes. However, the imputation accuracy was greater than in our study. Two-step imputation has proven to be more efficient and in our case could dramatically increase the imputation quality. Binsbergen et al. [30] obtained a correlation similar to ours (0.37) when imputing directly from 50 k to sequence level using 46 sequenced individuals. This correlation increased and reached 0.65 when an intermediate HD step was introduced. A HD chip is not yet available for caprine species, it would be a very powerful tool that would improve imputation to sequence level by overcoming imputation errors.

Our results varied depending on the imputation scenario used to impute the available 50 k genotypes to sequence level. In the Alpine breed, using a multi-breed reference panel resulted in the detection of a new signal for milk yield on chromosome 2 (Fig. 5). This signal does not appear when using the 50 k genotypes (Fig. 7). Significant variants in this region are annotated for CRYBA2, CDK5R2 and FEV genes which are not explicitly related to the mammary gland or any metabolic path linked to milk production. According to the RumimiR database, the region close to the signal is rich in miRNA: 12 miRNAs are located in a range of 1 Mb around the signal in the caprine species. However, among them, only 2 are expressed in mammary tissue: chr2_2187 (29.47 Mb) and chr2_2972 (29.80 Mb) and 4 are expressed in the ovaries: _Novel: bta-miR-153 (28.57 Mb), _Novel: bta-miR-26a (29.39 Mb), LO-m0073 (29.80 Mb) and FO-m0047 (29.80 Mb). Their exact role and impact on milk yield is still unknown. Further analysis is therefore required to confirm or disprove the signal detected on chromosome 2 for milk yield.

In the Saanen breed, the multi-breed reference panel led to the detection of a new signal for milk yield on chromosome 5 while confirming the involvement of a large area on chromosome 19. Significant variants on chromosome 5 are annotated for MDM1 gene, however the link between this gene and milk production is not clear. According to the RumimiR database, there are a few miRNAs in goats that are also located on chromosome 5 near our signal: novel_mir299 (44.53 Mb) discovered in blood samples, chr5_4536_mature (42.17 Mb) and chr5_4548_mature (45.82 Mb). The two latter are expressed at high levels in mammary tissue but are located further away from the signal and their exact involvement in milk yield is not known. As the signal only appeared when imputing using a French multi-breed reference panel, PCA was performed using PLINK for the region of chromosome 5 between 44.804 Mb and 44.816 Mb to try to understand where the imputed frequencies in the region came from. No significant breed group was found using PCA which implies that the QTL might actually be present in other French goat breeds.

In Saanen goats, a larger number of significant variants for milk production were detected on chromosome 19 and deeper investigation is required in this area that is also linked to udder health and conformation [6] and semen production. Multi-breed imputation gave the highest number of significant variants for milk yield (Table 4). The variants are annotated for 91 genes including 3 miRNAs (mir195, mir324 and mir497). Top 10 most significant variants (p-values between $2.96 * 10^{-14}$ and $1.09 * 10^{-15}$) are annotated for 2 genes (SCIMP and ZNF232). One of the 10 most significant variants (26,099,146) is situated in an intron of an unknown gene (GENE_id401516). Our study does not allow us to isolate functional candidate genes with certainty as it would require functional analysis. Nevertheless, the proximity of our signal to the ALOX genes cluster constitutes an interesting lead as the latter genes are implicated in lipid metabolism.

For semen production and more particularly semen volume, using a multi-breed reference panel considerably increased the noise observed on Manhattan plots (Fig. 6) making it difficult to distinguish true signals from what could be false positives. DYDs for this trait are more precise as they are derived from multiple repeated data from on average 100 daughters per buck whereas semen traits are the bucks' own limited number of repeated performances. The significance of the signal (Fig. 6) and the number of significant variants (Table 4) decreased slightly when a multi-breed reference panel was used compared with within-breed imputation. When imputing within-breed, 209 variants reached the chromosome significance level on chromosome 19. These variants are annotated for 61 genes. Four of the identified genes, (PELP1, ELP5, NEURL4 and CNTROB) are broadly expressed in testes. One gene (CHD3) is ubiquitous in the prostate, and another, YBOX2 (Y-box Binding Protein 2) is restrictedly expressed in testes. YBX2 is a member of the Y-box gene family that encodes a transcription factor and is specifically expressed in germ cells. Knock-out mice for this gene are of normal appearance but are sterile [33]. Mutations in this gene in humans are associated with male fertility disorders such as azoospermia and oligospermia [34]. A significant 23-bp deletion at position 26,614,373 was found in the French Saanen breed, close to the mature miRNA chi-miR-497 (26,614,406 – 26,614,427). The same

variant is also located near chi-miR-195 (26,614,085 – 26,614,104). Both miRNAs are ubiquitously expressed in testicular cells and might have an impact on semen production traits.

A pleiotropic region for milk, type traits and udder health was previously identified on chromosome 19 for the Saanen breed by Martin et al. [6]. Our study confirmed that a 3.5 Mb region was involved in milk production. For milk yield and semen volume, when sequences were imputed within-breed, top 10 variants had MAF comprised between 0.39 and 0.44 in the Saanen breed. The CLIP test rejected the pleiotropy assumption. The observed correlation was estimated at 0.013 and the threshold not to reject pleiotropy was above 0.15. The two traits might therefore be controlled by two different mutations situated close to each other. Moreover, none of the top 10 variants is shared between the two traits. According to the estimated effects, the allele with the highest frequency in the QTL region decreases both SV (− 0.09 SD) and milk yield (− 0.51 SD). Such an association therefore shows a favorable condition for improving both semen quantity and milk production.

Conclusions

This study provides insights on how to implement a robust quality check and an imputation pipeline based on caprine sequence data that will ensure the quality of subsequent analyses. New signals for milk yield traits were detected in both Alpine and Saanen breeds. Signals for semen and milk production traits were detected in the Saanen breed on chromosome 19. The latter regions however require further investigation and annotation to determine the genes involved and determine more precisely their impact. Imputation using a within-breed scenario appears to be more efficient because it is less time consuming. Signals detected after within-breed imputation show less noise and are more significant. However, due to the small size of our sequenced panel, within-breed imputation might not be able to detect smaller weaker signals. Increasing the number of sequenced animals should therefore be considered. Densifying the current genotyping array in the identified regions could corroborate their involvement in functional and production traits while removing potential imputation errors. In the same way, developing a HD chip for *Capra* species would improve the quality of imputation to sequence level by proceeding in two steps. Furthermore, functional analyses are required to confirm the involvement of identified genes in the studied phenotypes.

Abbreviations

AI: Artificial Insemination; ALOX: Arachidonate LipOXygenase; CDK5R2: Cyclin Dependent Kinase 5 Regulatory subunit 2; CHD3: Chromodomain Helicase DNA binding protein 3; CLIP test: Close Linkage versus Pleiotropy test; CNTROB: CeNTRoBin, Centriole Duplication and spindle assembly protein; CR: Concordance Rate; CRYBA2: Crystallin beta 2; DP: Depth; DYD: Daughter Yield Deviation; ELP5: Elongation acetyltransferase complex subunit 5; FEV: FEV transcription factor; GQ: Genotype Quality; GWAS: Genome Wide Association Study; HD: High Density; Indel: Insertion/Deletion; MAF: Minor Allele Frequency; MDM1: MDM1 nuclear protein; MiRNA: MicroRNA; MY: Milk Yield; NEURL4: Neuralized E3 ubiquitin protein ligase 4; PCA: Principal Component Analysis; PELP1: Proline, glutamate and leucine rich protein 1; QTL: Quantitative Trait Loci; R: Pearson Correlation; SC: Semen Concentration; SCIMP: SLP adaptor and CSK interacting membrane protein; SN: Spermatozoa Number; SNP: Single Nucleotide Polymorphism; SV: Semen Volume; WGS: Whole Genome Sequencing; YBOX2: Y-box Binding Protein 2; YD: Yield Deviation; ZNF232: Zinc Finger Protein 232

Acknowledgements

This study would not have been possible without the sequence data provided by the VarGoats Consortium (http://www.goatgenome.org/vargoats.html) and previous work by the International Goat Genome Consortium (IGGC, http://www.goatgenome.org/) and ADAPTmap Consortium (http://www.goatadaptmap.org/) providing relevant DNA samples, genotyping tools and genotyping data through their collaborative networks.

We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrénées and the CTIG (Centre de Traitement de l'Information Génétique) of INRAE Jouy-en-Josas for providing computing resources.

Discussions on sequence data quality and imputation methods with Mekki Boussaha (INRAE, UMR GABI) are gratefully acknowledged.

We also would like to thank the CapGenes breeding organization for the data provided.

We want to thank all VarGoats contributors: sample providers, technical staff from laboratories and sequencing platforms, system engineers, bioinformaticians and the VarGoats consortium steering committee. An updated and detailed list of people is available online.

Authors' contributions

CRG and RR designed the study. ET analyzed the data and drafted the manuscript. PB called the variants and provided support in computing. IP provided part of the performance file and chose individuals to be sequenced. CO and VC calculated the YDs for semen production traits. ET, GTK, CRG and RR interpreted the results. RR and CRG improved the manuscript. The VarGoats Consortium provided the sequence data. All authors read and approved the final manuscript.

Funding

The VarGoats project received financial support from France Génomique (ANR-10-INBS-09-08) through a call for Large Scale DNA Sequencing projects. The first author also received financial support from the Occitanie region and the Animal Genetics Division of the French National Institute for Agriculture, Food and Environment (INRAE-GA).

Availability of data and materials

The final sequence dataset will be made publicly available by the VarGoats Consortium. The use of the sequence data is under a data sharing agreement which is available here: http://www.goatgenome.org/vargoats_agreement.html and states that everyone will contact the VarGoats steering committee to discuss any publication plans that utilize this data to avoid the overlap of any planned analyses. Performance data and 50 k genotypes are not publicly available as they involve private professional partnerships.

Ethics approval and consent to participate

No animal experiments were conducted in this study, therefore no ethics approval was required.

Consent for publication

Not applicable.

Competing interests

The authors declare that they do not have any competing interests.

Author details
[1]GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan, France. [2]Sigenae, INRAE, 31326 Castanet-Tolosan, France. [3]Institut de l'Elevage, 31326 Castanet-Tolosan, France.

Received: 28 November 2019 Accepted: 13 February 2020

References

1. Rosen BD, Stella A, Rothschild MF, Tosser-Klopp G, Van Tassell CP, Crepaldi P, et al. AdaptMap: exploring goat diversity and adaptation. Genet Sel Evol. 2018;50(1):1–7.
2. Pausch H, Macleod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet Sel Evol. 2017; 49(1):1–30.
3. Bolormaa S, Chamberlain AJ, Khansefid M, Stothard P, Swan AA, Mason B, et al. Accuracy of imputation to whole-genome sequence in sheep. Genet Sel Evol. 2019;51(1):1.
4. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, et al. Correction: Design and Characterization of a 52K SNP Chip for Goats. PLOS ONE. 2016;11(3):e0152632.
5. Martin P, Palhière I, Maroteau C, Bardou P, Canale-Tabet K, Sarry J, et al. A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content. Sci Rep. 2017;7(1):1–13.
6. Martin P, Palhière I, Maroteau C, Clément V, David I, Klopp GT, et al. Genome-wide association mapping for type and mammary health traits in French dairy goats identifies a pleiotropic region on chromosome 19 in the Saanen breed. J Dairy Sci 2018;0(0):5214–5226.
7. Martin PM, Palhière I, Ricard A, Tosser-Klopp G, Rupp R. Genome wide association study identifies new loci associated with undesired coat color phenotypes in Saanen goats. PLoS One. 2016;11(3):1–15.
8. Martin P, Palhière I, Tosser-Klopp G, Rupp R. Corrigendum to "Heritability and genome-wide association mapping for supernumerary teats in French Alpine and Saanen dairy goats" (J. Dairy Sci. 99:8891–8900). J Dairy Sci. 2017; 100(9):7750.
9. Frischknecht M, Pausch H, Bapst B, Signer-Hasler H, Flury C, Garrick DJ, et al. Highly accurate sequence imputation enables precise QTL mapping in Brown Swiss cattle. BMC Genomics. 2017;18(1):1–10.
10. Sanchez MP, Govignon-Gion A, Croiseau P, Fritz S, Hozé C, Miranda G, et al. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet Sel Evol. 2017;49(1):1–16.
11. Huang S, He Y, Ye S, Wang J, Yuan X, Zhang H, et al. Genome-wide association study on chicken carcass traits using sequence data imputed from SNP array. J Appl Genet. 2018;59(3):335–44.
12. Nikbin S, Panandam JM, Yaakub H, Murugaiyah M, Sazili AQ. Novel SNPs in heat shock protein 70 gene and their association with sperm quality traits of Boer goats and Boer crosses. Anim Reprod Sci. 2014;146(3–4):176–81.
13. Mohammed KM, Khalil MH, Al-Saef AM. Genetic analysis for semen traits in a crossing program of Saudi Aradi with Damascus goats. Small Rumin Res. 2013;112(1–3):7–14.
14. Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb). 2014;112(1):39–47.
15. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49(4): 643–50.
16. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;00(00):1–3. https://arxiv.org/abs/1303.3997.
17. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303.
18. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program. SnpSift Front Genet. 2012;3.
19. Hickey JM, Kinghorn BP, Tier B, Van Der Werf JHJ, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol. 2012;44(1):1–11.
20. Sargolzaei M, et al. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15(1):478.
21. Johnston J, Kistemaker G, Sullivan PG. Comparison of different imputation methods. Interbull Bull. 2011;44(44).
22. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, et al. Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci. 2013;96(1):668–78.
23. Ventura RV, Miller SP, Dodds KG, Auvray B, Lee M, Bixley M, et al. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. Genet Sel Evol. 2016;48(1):1–20.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
25. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
26. Bourdon C, Bardou P, Aujean E, et al. RumimiR: a detailed microRNA database focused on ruminantspecies. Database. 2019;2019. https://doi.org/10.1093/database/baz099.
27. David I, Elsen J-M, Concordet D. CLIP test: a new fast, simple and powerful method to distinguish between linked or pleiotropic quantitative trait loci in linkage disequilibria analysis. Heredity (Edinb). 2013;110(3):232–8.
28. Li H, Sargolzaei M, Schenkel FS. Accuracy of whole-genome sequence genotype imputation in cattle breeds. 2014. https://doi.org/10.13140/2.1.2809.6642.
29. Ye S, Yuan X, Lin X, Gao N, Luo Y, Chen Z, et al. Imputation from SNP chip to sequence: a case study in a Chinese indigenous chicken population. J Anim Sci Biotechnol. 2018;9(1):1–12.
30. Van Binsbergen R, Bink MCAM, Calus MPL, Van Eeuwijk FA, Hayes BJ, Hulsegge I, et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2014;46(1):1–13.
31. Hancock DB, Levy JL, Gaddis NC, Bierut LJ, Saccone NL, Page GP, et al. Assessment of Genotype Imputation Performance Using 1000 Genomes in African American Studies. PLoS One. 2012;7(11).
32. Carillier C, Larroque H, Palhière I, Clément V, Rupp R, Robert-Granié C. A first step toward genomic selection in the multi-breed French dairy goat population. J Dairy Sci. 2013;96(11):7294–305.
33. Yang J, Medvedev S, Yu J, Tang LC, Agno JE, Matzuk MM, et al. Absence of the DNA–/RNA-binding protein MSY2 results in male and female infertility. Proc Natl Acad Sci. 2005;102(16):5755–60.
34. Hammoud S, Emery BR, Dunn D, Weiss RB, Carrell DT. Sequence alterations in the YBX2 gene are associated with male factor infertility. Fertil Steril. 2009;91(4):1090–5.

Publisher's Note

III.3. Analyses complémentaires : choix du logiciel d'imputation

Le choix des algorithmes et logiciels d'imputation s'est appuyé sur plusieurs arguments, le but principal étant d'adapter au mieux le procédé à la structure de nos données. Ainsi, nous avons tout d'abord envisagé une imputation familiale car elle a prouvé qu'elle était plus efficace et plus correcte qu'une imputation populationnelle quand le pedigree est disponible et de qualité suffisante (Antolín et al., 2017; Sargolzaei et al., 2014). En caprins laitiers en France, l'affiliation maternelle est systématiquement enregistrée et l'affiliation paternelle est toujours disponible quand le descendant est issu d'un mâle d'insémination artificielle. Nous avons également pris en compte les temps de calculs et choisi des logiciels pour lesquels ils étaient réduits. Pour cela de nombreuses études ont comparé l'efficacité des logiciels d'imputation (Boichard et al., 2012; Hickey, Kinghorn, Tier, Van Der Werf, & Cleveland, 2012; Johnston, Kistemaker, & Sullivan, 2011; Pausch et al., 2017; VanRaden et al., 2013; R. V. Ventura et al., 2014; Ricardo V. Ventura et al., 2016). Enfin nous avons essayé de prendre en compte l'incertitude d'imputation dans les analyses d'association en utilisant des dosages alléliques. Le dosage allélique est une variable comprise entre 0 et 2. Lorsque proche de 0 (respectivement 2) alors le génotype est plutôt homozygote d'un allèle (respectivement de l'autre). Une valeur proche de 1 nous indique que l'individu est probablement hétérozygote au locus.

En définitive, nous avons retenu le logiciel FImpute. En effet, dans de nombreuses études, il a prouvé son exactitude (VanRaden et al., 2013; R. V. Ventura et al., 2014; Ricardo V. Ventura et al., 2016). De plus, ce logiciel est très efficace du point de vue des temps de calcul (Johnston et al., 2011; VanRaden et al., 2013; Ricardo V. Ventura et al., 2016). Nous avons également comparé les sorties d'analyses d'association suite à l'imputation des séquences avec FImpute (génotype le plus probable) ou Minimac (dosage allélique). Minimac requiert un phasage préalable des données, il a été effectué à l'aide de MaCH (Fuchsberger et al., 2015). Les performances des deux logiciels ont été comparées sur le chromosome 6 connu pour son association au taux protéique (région des caséines) (Martin & Leroux, 2000; Teissier et al., 2018) et le chromosome19 porteur d'une région pléiotropique en Saanen (Martin et al., 2018). Sur la Figure 23 sont présentés les Manhattan plots sur le chromosome 19 pour les caractères de conformation de la mamelle, le score de cellules somatiques (LSCS), la production laitière et les caractères de production de semence. Comme le présente la Figure 23, les signaux sont complétement perdus en passant par Minimac. L'origine de cette perte n'est pas claire. Les sorties phasées de MaCH ont été comparées avec les entrées sans trouver

de modification de génotypes aux positions correspondant à des marqueurs de la puce. MaCH impute toutefois des génotypes aux positions pour lesquelles il est inconnu. En revanche, Minimac modifie entre 40,87 et 54,71% des typages de la puce 50k dans la population imputée alors que ces derniers sont connus et vérifiés au préalable. Ces modifications n'ont pas pu être expliquées, les génotypes n'étant pas retraités entre MaCH et Minimac, les

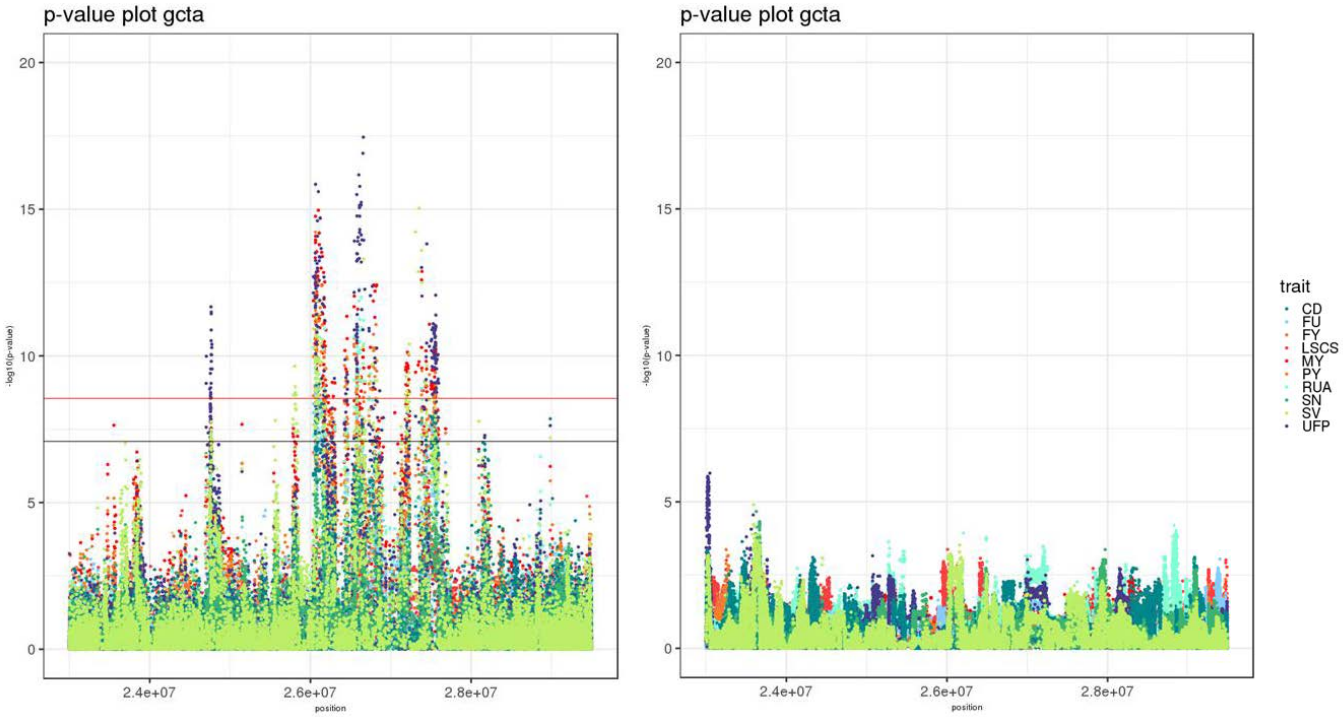


Figure 23: Manhattan plots sur le chromosome 19 pour des caractères identiques sur des séquences imputées par FImpute (à gauche) et Minimac (à droite)

modifications sont probablement liées au logiciel lui-même. Nous n'avons pas expérimenté ce dernier plus en détail.

tour de poitrine (CD), avant-pis (RU) qualité de l'attache arrière (RUA) et la position du plancher (UFP), santé de la mamelle : score de cellules somatiques (LSCS), production laitière : matière grasse (FY), matière protéique (PY), quantité de lait (MY) et enfin pour les caractères de production de semence : nombre de spermatozoïdes (SN) et volume de semence (SV).

IV. Conclusion du chapitre

Dans ce chapitre, nous avons détaillé les données de séquence caprines disponibles et leur composition tant en termes de races que de variants. Le dernier jeu de séquences du projet Vargoats représente 165 races dont une partie a été utilisée dans le cadre de cette thèse.

Le filtrage que nous avons implémenté s'appuie à la fois sur des filtres généraux qui utilisent des paramètres calculés par variant et des filtres qui ont été calibrés pour nos races

d'étude à l'aide des génotypages 50k disponibles. Cette méthode de filtrage s'est avérée efficace car nous avons obtenu des concordances 50k/séquence de 98,79% ±1,02 et 98,43% ±1,35 en Alpine et Saanen respectivement. Ce dernier a permis d'assurer la suite du traitement.

Ces séquences ont été mises à profit dans une étude prospective, la première en caprins, sur l'imputation. Différents scenarii d'imputation ont été testés et comparés. Une imputation intra-race permet d'obtenir une qualité des génotypes imputés correcte tout en limitant l'apparition de signaux parasites dans les analyses d'association ultérieures. La qualité d'imputation est encore problématique quand elle est comparée à des études similaires dans d'autres espèces. En effet, les corrélations entre génotypes imputés et génotypes vrais sont faibles. L'ajout de séquences en Alpines et Saanen françaises pourrait potentiellement répondre à ce problème.

Les analyses d'association sont impactées par le panel de référence utilisé pour l'imputation. Ainsi, de nouveaux signaux ont pu être détectés en utilisant les séquences de France métropolitaine. Ils nécessitent toutefois de plus amples investigations : ajout des variants significatifs sur une puce de génotypage ou analyses fonctionnelles pour être confirmés.

Le chromosome 19 en Saanen est impliqué dans plusieurs caractères de nature différente comme le confirme notre étude. En particulier, cette région influence des caractères de production laitière et de production de semence, ces derniers étant étudiés ici pour la première fois dans l'espèce caprine. Toutefois, l'analyse descriptive (significativité des SNP selon le caractère) et les tests statistiques (CLIP test) nous laissent à penser qu'il s'agit de plusieurs mutations proches sur le génome et non d'une mutation pléiotrope. Ce chromosome fera l'objet de plus amples investigations dans le chapitre suivant.

Chapitre 3
Approfondissement des données de séquence du chromosome 19 dans la race Saanen

Dans le chapitre précédent, nous avons observé un QTL sur le chromosome 19 dans la race Saanen pour la quantité de lait et le volume de semence. D'après de précédentes études sur génotypes 50k (P Martin et al., 2018), cette même région est associée à des caractères de production (matières protéique et grasse) et de santé de la mamelle (LSCS). Les caractères sont, à première vue, différents et indépendants, ce qui rend cette région complexe. Les tests de pléiotropie effectués sur génotypes 50k comme sur séquence ont rejeté l'hypothèse d'une seule mutation causale pour l'ensemble des caractères. Cependant, la région d'intérêt détectée est large (environ 5 Mb). Il reste donc difficile d'identifier un variant candidat pour chacun des caractères. Dans cette partie, nous exploiterons les données de séquence pour tenter d'affiner la localisation du QTL. Plusieurs approches ont été envisagées (ACP, analyses d'association, recherche de variants structuraux…) et s'appuient à la fois sur des données imputées et les données de séquences des seuls animaux séquencés. Ces travaux ont également contribué à la construction de la version 2 de la puce caprine d'Illumina par la sélection de variants dans la région QTL du chromosome 19. Cette partie comporte une *short communication* soumise le 04/07/2020 à *Genetics Selection Evolution*.

I. Utilisation des données de séquence pour la cartographie fine du QTL du chromosome 19 en race Saanen - Article

I.1. Introduction et résumé de l'article

La Saanen est une race caprine originaire de Suisse et plus précisément du Sud du Canton de Berne d'où elle tire son nom (Babo, 2000). Ses capacités de production sont reconnues et lui ont permis de conquérir rapidement de nouvelles régions du monde. D'après le système d'information sur les animaux domestiques de la FAO (http://www.fao.org/dad-is/transboundary-breed/en/; consulté le 12/05/2020), la Saanen est une des races les plus représentées à l'échelle mondiale. Elle est désormais officiellement présente dans 85 pays sur les 5 continents. Elle y est alors utilisée en race pure ou dans des schémas de croisement avec des races locales.

Nous l'avons vu dans le chapitre précédent, cette race est porteuse d'une importante région QTL sur le chromosome 19. Cette région couvre environ 5 Mb du chromosome et est à la fois associée à des caractères de production, conformation et stature de l'animal mais aussi à des caractères liés à la fertilité de la semence (P Martin et al., 2018; Talouarn et al., 2020). La multiplicité et la diversité des caractères impactés par cette région appellent un approfondissement de cette région du génome.

Une analyse ACP sur les données de séquences filtrées des races laitières européennes a été appliquée afin de mieux comprendre l'origine du QTL. Elle nous a permis d'établir que la Saanen française possède un profil particulier dans la région du QTL qui n'est pas retrouvé dans les autres races laitières d'Europe. En particulier, elle se différencie de la Saanen suisse. Différents profils génétiques ont pu être établis dans la région du QTL et sont associés à différents profils de performances pour les caractères associés au QTL. Ainsi, le profil qui présente les performances de production (lait MP, MG) les plus élevées possède également les moins bonnes performances en terme de conformation et santé de la mamelle. Ceci est cohérent avec les corrélations négatives observées entre ces caractères (Manfredi et al., 2001). Une recherche complémentaire sur les génotypes 50k nous a permis d'identifier 3 SNP de la puce qui permettent de différencier ces profils. Cette méthode nous a permis de confirmer la présence du QTL dans une race mixte néozélandaise et de supposer son absence dans la race Saanen canadienne.

I.2. La cartographie fine et la validation d'un QTL pléiotropique sur le chromosome 19 utilisant les données de séquence permet d'identifier trois profils phénotypiques et génotypiques chez les chèvres de race Saanen - Article

Fine mapping and validation of a pleiotropic QTL on chromosome 19 using sequence data identify three phenotypic and genomic profiles in Saanen goats

Estelle Talouarn[1,*], Gwenola Tosser-Klopp[1], Philippe Bardou[1,2], Virginie Clément[3], Isabelle Palhière[1], Hélène Larroque[1], Luiz F. Brito[4], Shannon Clarke[5], Christèle Robert-Granié[1], Rachel Rupp[1]

[1] GenPhySE, INRAE, Université de Toulouse, INPT, ENVT, 31326 Castanet Tolosan, France

[2] Sigenae, INRAE, 31326 Castanet-Tolosan, France

[3] Institut de l'Elevage, 31326 Castanet-Tolosan, France

[4] Department of Animal Science, Purdue University, West Lafayette, IN, 47907, United States of America

[5] AgResearch Limited, Invermay Agricultural Centre, Mosgiel, New Zealand

*Corresponding author

E-mail addresses:

ET: estelle.talouarn@inrae.fr; ORCID: 0000-0002-5016-0446

LB: britol@purdue.edu; ORCID: 0000-0002-5819-0922

SC: shannon.clarke@agresearch.co.nz; ORCID: 0000-0002-4615-8917

PB: philippe.bardou@inrae.fr; ORCID: 0000-0002-0004-0251

IP: isabelle.palhiere@inrae.fr

VC: virginie.clement@idele.fr

HL: Helene.larroque@inrae.fr

GTK: gwenola.tosser@inrae.fr; ORCID: 0000-0003-0550-4673

CRG: christele.robert-granie@inrae.fr; ORCID: 0000-0001-5313-2187

RR: rachel.rupp@inrae.fr; ORCID: 0000-0003-3375-5816

Abstract

Background

A wide QTL region located on chromosome 19 (CHI19) in Saanen goats has been previously reported to influence concomitantly production, udder health, type, and semen quality traits. However, no causal mutations nor candidate variants have been identified to date. Therefore, the main objectives of this study were to deepen our understanding of the QTL located on CHI19 based on whole-genome resequencing data from the VarGoats project, search for the QTL in other dairy goat populations, and provide supporting information to enable optimization of dairy goat breeding programs.

Results

We confirmed the association of an approximately 5-Mb region located on CHI19 with three production traits, four conformation traits, somatic cell score, and three semen production traits in dairy goats. The use of whole-genome sequence data enabled us to refine the QTL and define three genomic profiles in the region, but no candidate variants were identified. Indeed, we found a 3.6-Mb region in which both French and Italian Saanen goats clustered in three groups according to their genotypic profile while the Alpine breed animals were systematically clustered close to the Swiss Saanen and one of the homozygous group. The genotypic profiles are related to distinct multi-trait phenotypic performance and can also be distinguished using a combination of three SNPs from the Illumina GoatSNP50 BeadChip. Data from Canadian Saanen were similar to Alpine, suggesting no

segregation of this QTL in these other populations. On the other hand, data from New Zealand goats confirmed the presence of the QTL in the local mixed breed.

Conclusions

Our work represents an important step towards the understanding of the pleiotropic QTL region located on CHI19 in the Saanen breed. It provides unprecedented means for worldwide Saanen breeders to select breeding animals based on the desirable genomic profile. It is particularly interesting as it links production, conformation, and health traits that are usually evaluated separately.

Introduction

The Saanen goat breed originates from a Swiss breed from the Southern part of the Canton of Bern from which it takes its name [1]. Saanen goats have rapidly spread over the world, especially due to its high productivity and adaptability to different production systems and environmental conditions. According to the Food and Agriculture Organization – Domestic Animal Diversity Information System (www.fao.org/dad-is/transboundary-breed/en/), Saanen is one of the most represented dairy goat breeds worldwide. Saanen animals are raised as a pure breed or used in cross-breeding schemes with local breeds in at least 83 countries located on all 5 continents. In the early 2010s, a medium-density genotyping chip (50K, Illumina GoatSNP50 BeadChip; [2]) was developed by the International Goat Genome Consortium (IGGC) [2] and is being routinely used by the goat industry. In the Saanen breed, several genome-wide association analyses using this tool revealed the implication of an approximately 5-Mb region located on chromosome 19 (CHI19) for udder health and type traits [3–5]. The same QTL region was concomitantly observed in a Scottish mixed-breed population (composited breed developed by crossing Alpine, Saanen, and Toggenburg animals) for conformation traits and milk yield [6]. Sequence data revealed the association of milk yield and semen volume to the same region of the genome [7]. The QTL seems to be breed-specific and was not

72 observed in French Alpine, even though the Alpine and Saanen breeds have genetically diverged from
73 few generations.

74 The region explains between 5 and 10% of the total additive genetic variance of somatic cell
75 score (SCS) and type traits [5]. Alleles from significant SNPs, which have a positive effect in milk,
76 fat, and protein yields, and semen volume, tend to deteriorate udder type and health traits. This
77 conflict is in accordance with the negative correlation that has been revealed in Saanen goats between
78 milk production and udder type traits [8] and between milk production and LSCS (Lactation average
79 SCS) [9]. However, it is unknown whether the same mutation is associated to all traits mentioned or
80 if there are multiple causal mutations located nearby the region identified [5,7]. Only the latter option
81 can allow disrupting the pleiotropic effects of the region. Deeper investigation is therefore needed in
82 order to better understand this region of the genome.

83 The recent incoming of whole-genome resequence data (VarGoats project;
84 www.goatgenome.org/vargoats.html) is a great opportunity to unravel the complexity of the region
85 as it covers more finely the genome than the Illumina GoatSNP50 BeadChip [2]. Indeed, whole-
86 genome sequence data comprise variants that were not selected to be genotyped with the Illumina
87 GoatSNP50 BeadChip, including low Minor Allele Frequency (MAF) markers, small
88 insertions/deletions (indels), variants poorly mapped on the previous version of the genome
89 (CHIR1.0), and variants not detected in previous whole-genome sequence datasets. In this context,
90 the main objectives of this study were to deepen our understanding of the QTL located on CHI19
91 based on whole-genome resequencing data from the VarGoats project, including fine mapping on the
92 multi-trait dimension, validation in other dairy goat populations; and to provide supporting
93 information to enable optimization of dairy goat breeding programs.

Methods

Whole-genome sequence data

No Animal Care Committee approval was necessary for the purposes of this study, as all information required was obtained from pre-existing databases. The complete description of the datasets used is provided in Talouarn *et al.* [7]. The sequence data were acquired from the VarGoats project (www.goatgenome.org/vargoats.html). We used whole-genome sequence data from 829 goats (*Capra hircus*) of various breeds and geographical origins. We removed 14 individuals, including four French Saanen, with a mean sequencing coverage depth lower than 5 to reduce genotype uncertainty. Thus, the final dataset consisted of 813 sequenced individuals and comprised 33 French Saanen (31 males and 2 females), 4 Swiss Saanen and 5 Italian Saanen.

The sequence quality control and filtering is described in details by Talouarn *et al.* [7]. In brief, 23,337,436 SNP variants, including 508,023 variants on CHI19 remained for further analyses. Concordance with the 50K genotypes was checked on the overall genome. After variant filtering, the individual mean concordance rate was 98.24% and ranged from 94.00% to 99.96% for both French Alpine and Saanen. A within-breed genotype imputation was necessary to fill in the gaps. AlphaImpute (v 1.9) [10] and FImpute (v 3.0) [11] were therefore used consecutively. After imputation, the mean concordance rate between 50K genotypes and whole-genome sequence data in the CHI19 chromosome in both French Alpine and Saanen was 98.43% (± 1.35). There were no missing genotypes after the imputation analysis. For subsequent analyses, as 50K genotypes are more reliable than low-depth sequencing, and to avoid spreading genotyping errors down the pedigree, the common SNPs (n=40,491) between the 50K SNP panel and whole-genome sequence data were systematically replaced by the 50K genotypes.

Genotypes with the GoatSNP50 BeadChip and imputation to sequence level

A total of 757 French Saanen bucks were genotyped with the Illumina GoatSNP50 BeadChip [2]. Prior to the genotype imputation, a genotype quality control was performed as described in Martin *et al.* [5]. In brief, individuals with call rate lower than 95% or showing pedigree inconsistency were discarded. SNP quality control was based on the following inclusion criteria: (1) call rate greater than 99%, (2) minor allele frequency above 1%, and (3) Hardy-Weinberg Equilibrium P-value greater than 10^{-6}. A total of 47,147 out of the 53,347 SNPs remained on goat autosomes, including 772 SNPs on CHI19 that were used for further analyses.

For the association analyses, all genotyped males were then imputed to whole-genome sequence level using filtered sequence data. Imputation was performed within-breed using FImpute (v 3.0) software [11]. Pedigree information was available and used as the genotyped individuals were closely related to the reference panel of sequences. Mean genotype and allele concordance rates were 0.76 and 0.87 in Alpine, respectively, and 0.74 and 0.86 in Saanen. Furthermore, we used 50K genotypes of 831 Canadian Saanen and 793 Canadian Alpine. The data collection and filtering has been described in Brito et al. [12]. We also included 52 genotypes (50K SNP panel) of mixed breed (Saanen and Swiss Alpine) from New Zealand.

Phenotypes

A total of 522 males with phenotypes were used in this study. The description of the male (semen) and female (milk production and composition, health, and type or conformation) traits is presented in Talouarn *et al.* [7] and summarized hereafter.

142

143 *Male traits*

144

145 Three semen production traits were studied on 471 artificial insemination (AI) bucks: semen
146 volume in mL (SV), semen concentration in billions of spermatozoa per mL (SC) and number of
147 spermatozoa in billion spermatozoa (SN). As these records can be directly measured on the males,
148 yield deviations (YD) were used as the phenotypes for the 471 AI bucks for which semen production
149 traits have been measured.

150

151 *Female traits*

152

153 Traditional milk production traits were analyzed: milk yield (MY; kg), protein yield (PY; kg),
154 fat yield (FY; kg), and LSCS. Four type (or body conformation) traits were considered for this study,
155 including fore udder (FU), udder floor position (UFP), rear udder attachment (RUA), and chest depth
156 (CD). All type trait measurements and LSCS have been described in details by Martin *et al.* [3,5].
157 For all the female traits, the Daughter YD (DYD) were computed for the 522 bucks with at least 30
158 daughters with records. DYD were the average daughters' performance corrected for environmental
159 effects and genetic merit of the dam.

160

161 ## Principal Components Analyses on filtered sequence data

162

163 Principal Component Analysis (PCA) were conducted using the PLINK (v1.9) software [13]
164 and filtered "vcf" data files. The Saanen breed was compared to various panels of breeds from the
165 813 individuals sequenced worldwide were used. PCA was performed on variants of either the QTL
166 region of CHI19 (between 23 and 30 Mb) or on filtered data of CHI1 of all European *Capra hircus*

individuals. This was done to compare the results as CHI1 is not known to harbor any QTL locus. Variants with MAF lower than 0.01 were filtered out. With these analyses, we aimed to validate whether the QTL segregating in French Saanen is also segregating in other Saanen populations or other dairy goat breeds, and thus, this would contribute to tracing back this QTL origin.

Association analyses on imputed sequence data

The imputed sequences of CHI19 were subjected to within-breed and single-trait association analysis for the 11 traits related to milk production, udder type, and semen production in French Saanen bucks. We used the following mixed linear model with the *lmm* option implemented in the Genome-wide Efficient Mixed Model Association (GEMMA) software [14] :

$$\boldsymbol{y} = 1\mu + \boldsymbol{x}b + \boldsymbol{u} + \boldsymbol{e}$$

where **y** represents the pre-adjusted phenotypes of each trait; μ is the overall mean; b is the additive fixed effect of the variant tested; $\boldsymbol{x}$ is the vector of imputed genotypes coded as 0, 1, 2 (number of copies of the alternative allele); $\boldsymbol{u}$ is the vector of random additive polygenic effects, $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{G}\sigma^2)$ with **G** the genomic relationship matrix; $\boldsymbol{e}$ is the vector of normally distributed random residual effects. The genomic relationship matrix (**G**) was calculated based on the 50K genotypes and using the GEMMA software and the parameter "*gk*". Variants with a within-breed MAF lower than 0.01 were excluded, leaving 272,835 variants on CHI19 in French Saanen. A Bonferroni correction (6.74) was applied to account for multiple testing.

Results and discussion

Association analyses

191 The eleven traits (MY, PY, FY, CD, RUA, UFP, LSCS, SN, SC, and SV) were found to be
192 significant on an approximately 5.2-Mb region of CHI19 in French Saanen (Figure 1). A total of 784
193 variants were found significant for at least one trait on CHI19 including 34 50K SNPs. Details on the
194 significant variants and significance level reached per trait are presented on Table 1 and Figure 1.
195 Significant variants were annotated for 96 genes with an identified function, 16 tRNA, and three
196 miRNA (Additional File 1). The variety of traits associated with this genomic region rises several
197 questions especially as signals are not trait-specific. Candidate variants were not identified based on
198 the association analysis. However, the significant variants close to miRNA might be interesting key
199 variants to investigate as their target mRNA might be linked to the traits we are studying.
200
201 **Table 1. Summary of the significant variants identified on the 5.2-Mb region of chromosome**
202 **19 of Saanen goats using imputed whole-genome sequence data.**

Trait	Number of significant variants	Number of significant GoatSNP50 BeadChip SNP variants	Lowest p-value	Position of the most significant variant	MAF of the most significant variant
CD	100	10	2.84×10^{-11}	26,824,298	0.42
FU	100	7	1.11×10^{-11}	26,148,755	0.41
FY	174	9	4.82×10^{-12}	27,506,923	0.29
LSCS	34	1	7.90×10^{-10}	28,043,599	0.50
MY	455	21	1.61×10^{-17}	26,653,421	0.43
PY	322	16	1.08×10^{-15}	26,653,421	0.43
RUA	293	11	8.16×10^{-16}	26,662,281	0.38
SC	16	1	4.59×10^{-9}	24,692,223	0.35
SN	9	1	9.48×10^{-9}	26,662,281	0.39

SV	361	20	2.53×10^{-15}	26,746,150	0.45
UFP	637	32	3.69×10^{-20}	26,653,421	0.43

CD: chest depth, FU: fore udder, FY: fat yield, LSCS: average lactation somatic cell score, MY: milk yield, PY: protein yield, RUA: rear udder attachment, SC: semen concentration, SN: number of spermatozoa, SV: semen volume, UFP: udder floor position.

Please insert Figure 1 here

PCA on sequence data

The CHI19 QTL region was confirmed in the French Saanen breed but was not found to be associated with any trait evaluated in the French Alpine breed [7]. A PCA using sequence data variants located in the QTL region of CHI19 was therefore performed to compare the genomic background of the two breeds to better understand the architecture of the QTL region. The result is presented on Figure 2. The first two axes of the PCA account for ~25% of the total variation. Saanen individuals clustered together with French Alpine animals and another group of Saanen animals clustered further away on the first principal component. As the QTL signal detected in this region does not appear in the Alpine breed, it seemed interesting to explore the difference between these groups. This was done in two-steps, after defining three Saanen groups: Group 1: animals with a coordinate lower than -0.15 on the first principal component (PC1; Figure 2), Group 3: included animals with a positive coordinate on PC1, and, Group 2: included all remaining intermediate individuals. First, we investigated the Daughter YD of individuals of each group for all 11 traits. We then compared means of each Saanen group based on a Student's t-test. Mean sequencing depth and birth year were also analyzed but no statistically significant difference ($P > 0.05$) was found for both variables between groups. Second, aside from a simple comparison between breeds, a Component

Analysis was performed using the "*CA*" function from the "FactoMineR" R package on French Alpine and Saanen of groups 1 and 3 including variants from the QTL region. This analysis enabled us to extract variants that were differentiating the groups on the CA graphs to look closer at their genotypes for each individual. We aimed at defining a group of variants that would differentiate the Saanen animals of Group 1 from Alpine and Saanen of Group 3. Only 2,545 variants with a coordinate lower than -1 on PC1 were retained to analyze the genotypes of each individual. The genotypes for these variants were extracted from the filtered whole-genome sequence data. We found a 3.6 Mb region in which Alpine and Saanen individuals that clustered close to Alpine are homozygote for one allele and the Saanen animals that clustered away from the Alpine animals were homozygous for the other allele. Most of the Saanen individuals between these two groups are heterozygote. Based on these results, we identified an interesting group of variants that might be explaining the width of the signal detected for various economically-important traits in French Saanen. However, the width of the haplotype did not enable us to point out specific genes implicated in the metabolic pathways underlying the phenotypic expression of the 11 traits investigated. Therefore, further functional analyses are recommended for that purpose. Nonetheless, genotypes for this group of variants enable the distinction of phenotypic profiles of Saanen animals (Table 2). However, due to the small number of individuals per group, the comparison of groups was only significant for UFP, FU, and SC. Group 1 that clusters the furthest away from the French Alpine breed tend to have higher levels of milk production and composition as well as higher semen quality, but they also show deteriorated udder type and health traits. This is consistent with the negative correlation found by Manfredi *et al* [8]. Group 2 is an interesting intermediate group with good milk production and intermediate performances for all traits. Group 3 animals have lower production performances, but improved type traits and semen concentration. In this context, it might be interesting for dairy goat breeding organizations, such as Capgenes (www.capgenes.com), to identify crossovers within the haplotype

that could enable selection for greater milk and semen production traits while still improving or maintaining udder type and health traits.

We also investigated if the differentiation on CHI19 was also present in other French or European dairy goat breeds. There are four groups in the French dairy goat breeds based on CHI1: (1) Saanen, (2) Alpine and Savoie, (3) Lorraine and Poitevine, (4) Fossés, Provençale, Pyrénées, and Rove (Figure 3A). Saanen then tend to cluster away from the other French breeds. This difference is heightened on the QTL region of CHI19 as the percentage of variance explained by PC1 reaches 16.8% (Figure 3B). Interestingly, on CHI19 between 23 and 30 Mb, when analyzing all European goats with whole-genome sequence data available, the same pattern can be observed (Figure 4). The French and Italian Saanen sequenced individuals clustered away from breeds from other countries. They define the first component and the Swiss Saanen does not segregate with this group. Surprisingly, the Swiss Saanen lies closer to the Alpine breed. French Alpine and Saanen breeds originate from the same ancestral breed from Switzerland [1] and diverged very recently. Our results suggest that French and Swiss Saanen are already genetically drifting apart. However, this has to be put in perspective with the very few number of Swiss Saanen sequenced individuals (only 4 in our dataset). They might actually not represent the genetic architecture of the Swiss Saanen population. However, Italian and French Saanen are genetically similar based on this QTL region. This can be explained by the frequent importation of semen from French Saanen AI bucks, which is erasing differences between countries (Isabelle Palhière, INRAE, personal communication). Furthermore, similar breeding goals might also contribute to this genetic similarity between these two Saanen populations.

Please insert Figures 2 to 4 here

Please insert Table 2 here

Searching for 50K SNP variants to distinguish Saanen profiles

In the light of our findings on sequence variants, we searched for 50K markers that would differentiate Saanen profiles. This was done to provide a simple decision tools for individual selection of breeding animals. Fifty-seven markers of the 50K genotypes were located on the QTL region of CHI19 were extracted for the 33 Saanen and 40 Alpine individuals in an attempt to investigate the possibility of attributing a group to a Saanen that would only be genotyped with the Illumina GoatSNP50 BeadChip. We performed a *pls-da* analysis using the *mixOmics* R package, based on the 50k genotypes of the sequenced French Alpine and Saanen. Groups were assigned based on the PCA results (Figure 2). This analysis allowed us to rank the SNPs according to their ability to distinguish groups (weight on PC1). We then looked for markers with the following characteristics (1) homozygote for an allele in the 40 Alpine and the six Saanen individuals of Group 3 (2) heterozygote for the 17 Saanen individuals of Group 2, and (3) homozygote for the second allele for the 10 Saanen individuals of Group 3. We retained variants according to (1) their ranking, (2) their sensibility and specificity, and (3) the lowest number of missing genotypes in available 50K genotypes. We also checked that the three markers retained were able to maximize the number of genotyped Saanen within a group. Genotypes for the three SNPs for the three groups are presented in Table 3. Two SNPs were found significant in the association analyses (Table 1), in which the first SNP (26,148,755) for FU and the second (26,662,281) for RUA and SN. Specificity and sensibility values of the group attribution using these three variants are presented in Table 4. We considered that an individual was part of one of the group if it had the correct genotype for at least two of the three SNP variants. Results of groups attribution are presented in Table 5. This method accurately defined a profiling group for 534 out of the 541 Saanen males with both genotypes and phenotypes. The method identified only 10 Alpine individuals out of 2,409 as part of Group 1 which represents only 0.42% of available

299 genotypes. According to the Hardy Weinberg Equilibrium principle, if we consider that Group 1 and
300 Group 3 are homozygote and Group 2 is heterozygote, the number of individuals per group would
301 imply a MAF comprised between 35 and 45%. This expectation is consistent with the frequency
302 observed for the most significant variants found in the QTL region (Table 1).

303 Among the 534 males, 396 had performance records for the traits associated with the CHI19
304 QTL. The differences between groups were significant ($P < 0.05$) for all traits, except for SV and
305 LSCS in which Group 3 was not significantly different of Group 2. Figure 5 shows a radian plot of
306 the mean profile observed for each group of Saanen for the 11 traits. The same trend observed on
307 sequenced individuals was also observed in the genotyped males. Indeed, Saanen from Group 1 tend
308 to have higher levels of milk production and composition, but also higher SCS and poorer udder type.
309 Means and standard deviations for each trait associated with the CHI19 QTL are presented in
310 Additional File 2.

311 With the aim of applying our classification to other worldwide Saanen populations, we
312 evaluated our method in other dairy goat breeds, including Saanen, from Italy and Switzerland. We
313 extracted the three SNPs from the sequence data file and attributed groups to 699 out of the 813
314 sequenced individuals (Additional Table 3). When excluding Saanen individuals from the analysis,
315 only 9.7% of the animals were attributed to Group 1, while 37.4% of individuals genotyped with the
316 Illumina GoatSNP50 BeadChip in the French Saanen population were attributed to Group 1. In the
317 Swiss Saanen, no individual was found in Group 1, which is consistent with the PCA results. Out of
318 the five Italian Saanen, we found two individuals in Group 1 and three in Group 2, which confirms
319 that the Italian Saanen also carries the same genomic profile than the French Saanen in this specific
320 region of the genome. Therefore, the three selected SNPs can be used to differentiate the groups. This
321 is a promising perspective for Saanen breeders as the Illumina GoatSNP50 BeadChip is already
322 widely used across the world.

Subsequently, we evaluated the genomic profiling of Canadian Saanen and Alpine as well as mixed breed individuals from New Zealand, based on 50K genotypes, to check if the method of assignment was consistent across worldwide Saanen populations (Table 5). Interestingly, the frequency of Group 1 in Canadian Saanen is closer to the observed frequency in both French and Canadian Alpine than in French Saanen. This suggests that this QTL is not segregating in the Canadian Saanen population. These results are consistent with the development history of Canadian dairy goat populations [12,15]. The proximity between Canadian Saanen and Alpine is likely due to a greater level of crossbreeding between Alpine and Saanen as well as some contributions from other dairy goat breeds (e.g., Toggenburg and LaMancha). Despite French and Canadian Saanen being genetically related and considered as the same breed, the selection intensity as well as traits included in the breeding program are different in both countries. In the New Zealand mixed breed population, the CHI19 QTL region has been maintained at frequencies close to those found in French Saanen, suggesting that this part of the genome has been inherited from the French Saanen ancestry in the mixed breed population. These results confirm the ability of our method to find the CHI19 QTL in other goat populations. Interestingly, a QTL similar to the one observed in French Saanen is indeed present in New Zealand dairy goats between 24,836,694 and 28,953,102 on CHI19 for MY, PY, FY, and LSCS [16].

Table 3. Genotype for each of the three markers selected from the Illumina GoatSNP50 BeadChip for genomic group profiling

	Genotype at marker 1	Genotype at marker 2	Genotype at marker 3
Position (bp)	26,148,755	26,662,281	28,169,892
Group 1	AA	GG	GG
Group 2	AG	AG	AG
Group 3	GG	AA	AA

Table 4. Specificity and sensibility of the Saanen profile attribution using 3 SNP chip markers for group attribution

Group	Marker	1	2	3
1	Sensibility	100.0%	100.0%	80.0%
	Specificity	94.8%	93.1%	91.4%
2	Sensibility	92.9%	78.6%	78.6%
	Specificity	72.2%	87.0%	63.0%
3	Sensibility	61.4%	81.8%	54.6%
	Specificity	100.0%	100.0%	100.0%

Table 5. Results of profile assignment using 3 SNPs of 50k genotypes.

Country of origin	Breed	Group 1 (%)	Group 2 (%)	Group 3 (%)
France	Saanen	37.4	42.7	19.9
	Alpine	0.4	23.1	76.5
Canada	Saanen	2.0	21.8	76.2
	Alpine	0.4	26.0	73.6
New Zealand	Mixed	35.6	35.6	28.9

Please insert Figure 5 here

Conclusions

This study provides an in-depth description of a QTL region located on chromosome 19, based on whole-genome sequence information from a variety of dairy goat breeds. The causal mutations were not unraveled, but we provided the scientific knowledge to develop a global profiling of dairy goats based in only three SNPs included in the Illumina GoatSNP50 BeadChip, which is a genotyping platform used worldwide. This genomic region is of great importance as it links production (milk and semen), type or conformation, and health traits, which are usually evaluated separately.

List of abbreviations

AI: Artificial Insemination

CD: chest depth

IGGC: International Goat Genome Consortium

FU: fore udder

FY: fat yield

LSCS: somatic cell score

MY: milk yield

PCA: Principal Component Analysis

PY: Protein Yield

RUA: rear udder attachment

SC: semen concentration

SN: number of spermatozoa

SV: semen volume

UFP: udder floor position

Declarations

Ethics approval and consent to participate

No Animal Care Committee approval was necessary for the purposes of this study, as all information required was obtained from pre-existing databases.

Consent for publication

Not applicable

Availability of data and materials

The final sequence dataset will be made publicly available by the VarGoats Consortium. Performance data and 50K genotypes are not publicly available as they involve private professional partnerships.

Competing interests

The authors declare that they do not have any competing interests.

Funding

The VarGoats project received financial support from France Génomique (ANR-10-INBS-09-08) through a call for Large-scale DNA Sequencing projects. The first author also received financial support from the Occitanie region and the Animal Genetics Division of the French National Institute for Agricultural Research (INRAE-GA). This research used genotyping data from projects funded by the sector councils of Quebec, Ontario and British Columbia, who administer the Canadian Agricultural Adaptation Program (CAAP) for Agriculture and Agri-Food Canada; Ontario Goat; Société des éleveurs de chèvres laitières de race du Quebec; GoatGenetics.Ca; and the Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA), through the Ontario Agri-Food Innovation Alliance.

Authors' contributions

CRG and RR designed the study. ET analyzed the data and drafted the manuscript. PB called the variants and provided computational support. IP, HR, and VC provided part of the performance file and chose the individuals to be sequenced. VC calculated the yield deviations for the semen production traits. SC provided 50K genotypes from New Zealand. ET, GTK, CRG, LFB, and RR

interpreted the results. RR, LFB, and CRG edited and improved the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study would not have been possible without the sequence data provided by the VarGoats Consortium (http://www.goatgenome.org/vargoats.html) and previous work by the International Goat Genome Consortium (IGGC, www.goatgenome.org/) and ADAPTmap Consortium (www.goatadaptmap.org/) providing relevant DNA samples, genotyping tools and genotyping data through their collaborative networks. We are also grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees and the CTIG (Centre de Traitement de l'Information Génétique) of INRAE Jouy-en-Josas for providing computing resources. We also would like to thank the Capgenes breeding organization for the data provided. We are grateful to Hayley Baird from AgResearch Animal Genomics for providing access to the NZ goat DNA samples.

References

1. Babo D. Races ovines et caprines françaises. Editions F. 2000.
2. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al. Design and characterization of a 52K SNP chip for goats. PLoS One. 2014;9(1).
3. Martin P, Palhière I, Maroteau C, Bardou P, Canale-Tabet K, Sarry J, et al. A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content. Sci Rep. 2017;7(1):1–13.
4. Martin PM, Palhière I, Ricard A, Tosser-Klopp G, Rupp R. Genome wide association study identifies new loci associated with undesired coat color phenotypes in Saanen goats. PLoS One. 2016;11(3):1–15.

5. Martin P, Palhière I, Maroteau C, Clément V, David I, Tosser-Klopp G, et al. Genome-wide association mapping for type and mammary health traits in French dairy goats identifies a pleiotropic region on chromosome 19 in the Saanen breed. J Dairy Sci. 2018;0(0):5214–26.
6. Mucha S, Mrode R, Coffey M, Kizilaslan M, Desire S, Conington J. Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. J Dairy Sci. 2017;101(3):2213–25.
7. Talouarn E, Bardou P, Palhière I, Oget C, Clément V, The VarGoats Consortium, et al. Genome wide association analysis on semen volume and milk yield using different strategies of imputation to whole genome sequence in French dairy goats. BMC Genet. 2020;21(1):1–13.
8. Manfredi E, Piacere A, Lahaye P, Ducrocq V. Genetic parameters of type appraisal in Saanen and Alpine goats. Livest Prod Sci. 2001;70(3):183–9.
9. Rupp R, Clément V, Piacere A, Robert-Granié C, Manfredi E. Genetic parameters for milk somatic cell score and relationship with production and udder type traits in dairy Alpine and Saanen primiparous goats. J Dairy Sci. 2011 Jul;94(7):3629–34.
10. Hickey JM, Kinghorn BP, Tier B, Van Der Werf JHJ, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol. 2012;44(1):1–11.
11. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15(1).
12. Brito LF, Kijas JW, Ventura R V., Sargolzaei M, Porto-Neto LR, Cánovas A, et al. Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. BMC Genomics. 2017;18(1):1–20.
13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a

tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

14. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014 Feb 16;11:407.

15. Grossi DA, Jafarikia M, Brito LF, Buzanskas ME, Sargolzaei M, Schenkel FS. Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in Canadian pigs. BMC Genet. 2017;18(1):1–13.

16. Scholtens M, Jiang A, Smith A, Littlejohn M, Lehnert K, Snell R, et al. Genome-wide association studies of lactation yields of milk, fat, protein and somatic cell score in New Zealand dairy goats. J Anim Sci Biotechnol. 2020;11(1):1–14.

Figures

Figure 1 Manhattan plot on the QTL region located in chromosome 19 for three milk production and composition traits, three semen traits, lactation average somatic cell score, and four type traits.

CD: chest depth, FU: fore udder, FY: fat yield, LSCS: lactation average somatic cell score, MY: milk yield, PY: Protein Yield, RUA: rear udder attachment, SC: semen concentration, SN: number of spermatozoa, SV: semen volume, UFP: udder floor position.

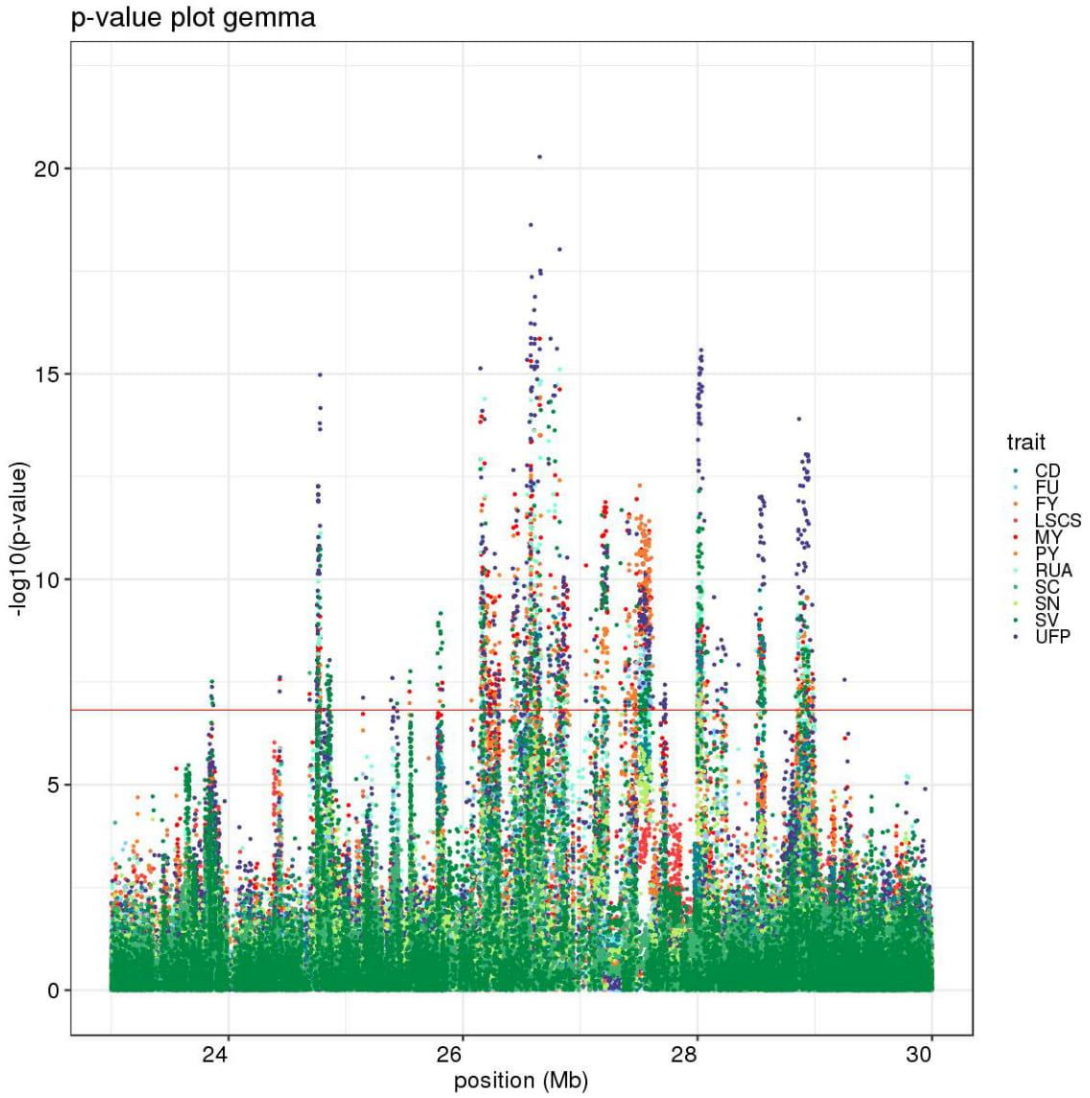


Figure 2 Principal Component Analysis performed based on sequence data variants located in a QTL region (23 to 30 Mb) in chromosome 19 for 33 French Saanen and 40 French Alpine animals.

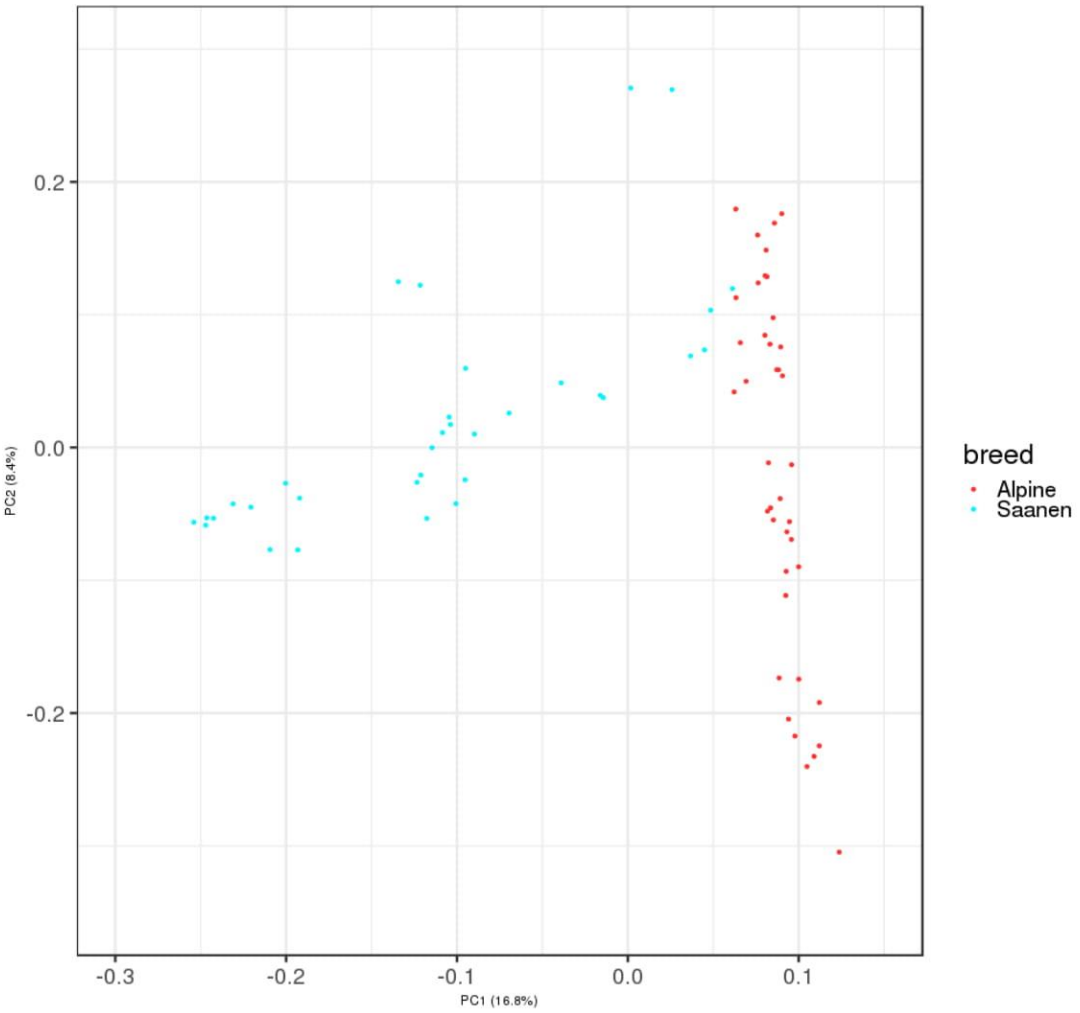


Figure 3 Principal Component Analysis performed based on sequence data variants located in the chromosome 1 (A) and chromosome 19 (B) for 170 French dairy goats with whole-genome sequence data.

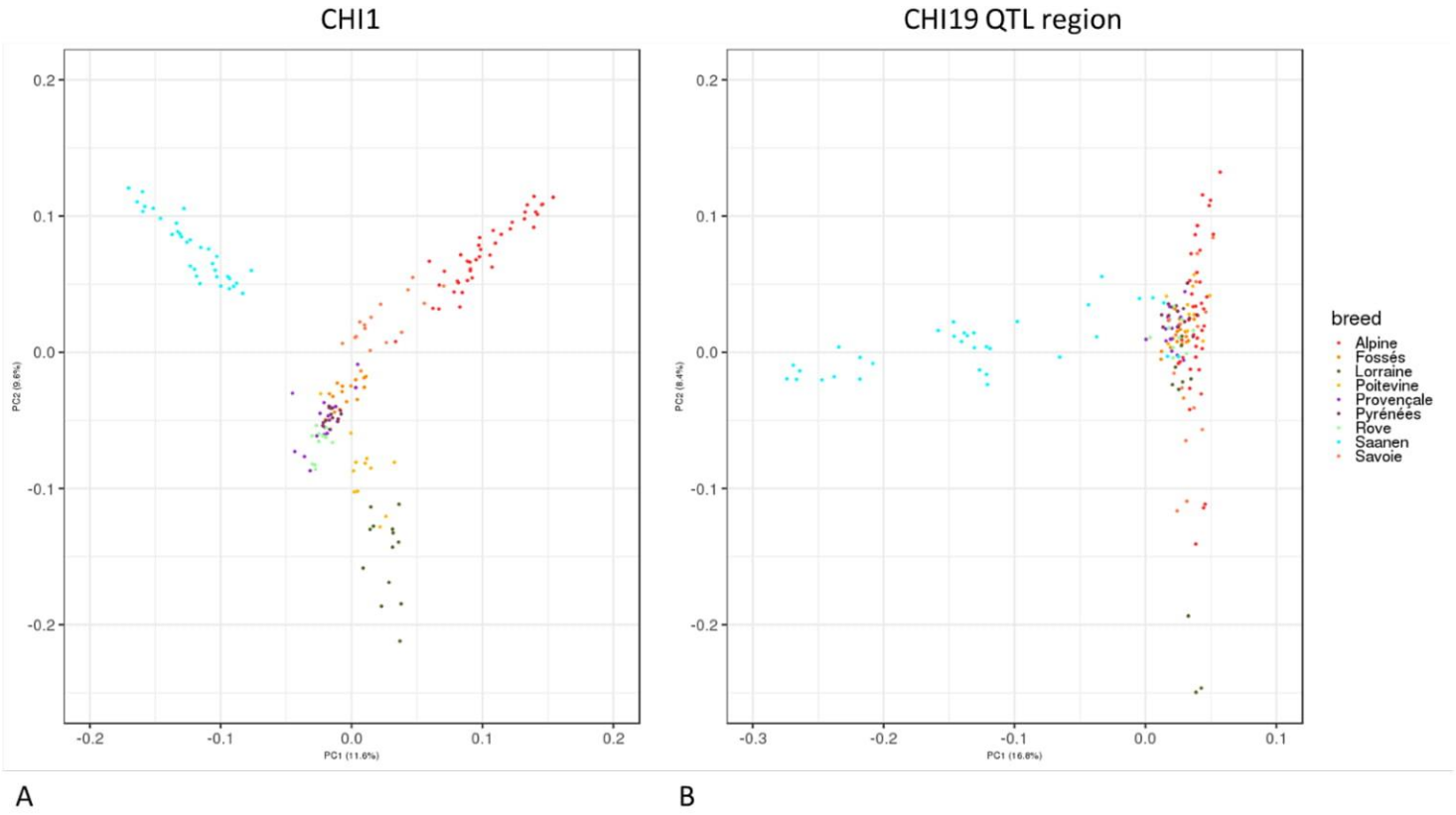


Figure 4 Principal Component Analysis performed based on sequence data variants of QTL region of chromosome 19 (23 to 30 Mb) for 244 European dairy goats with whole-genome sequence data.

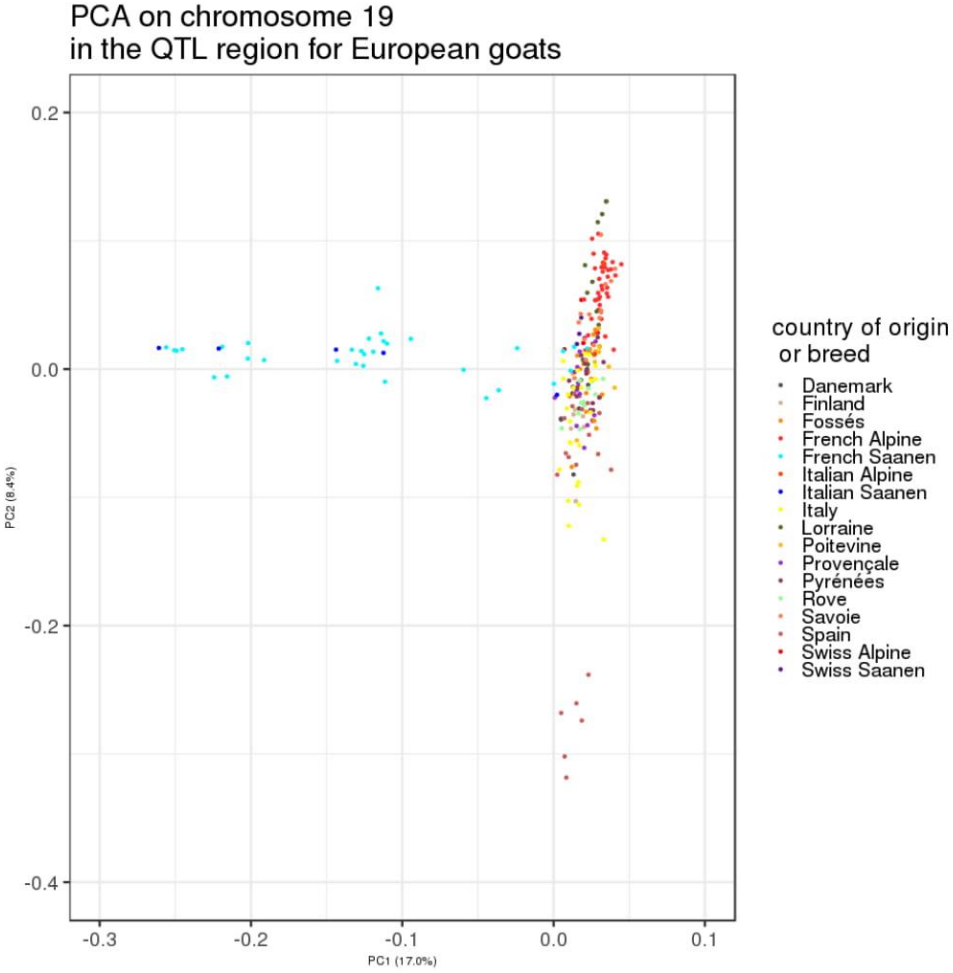
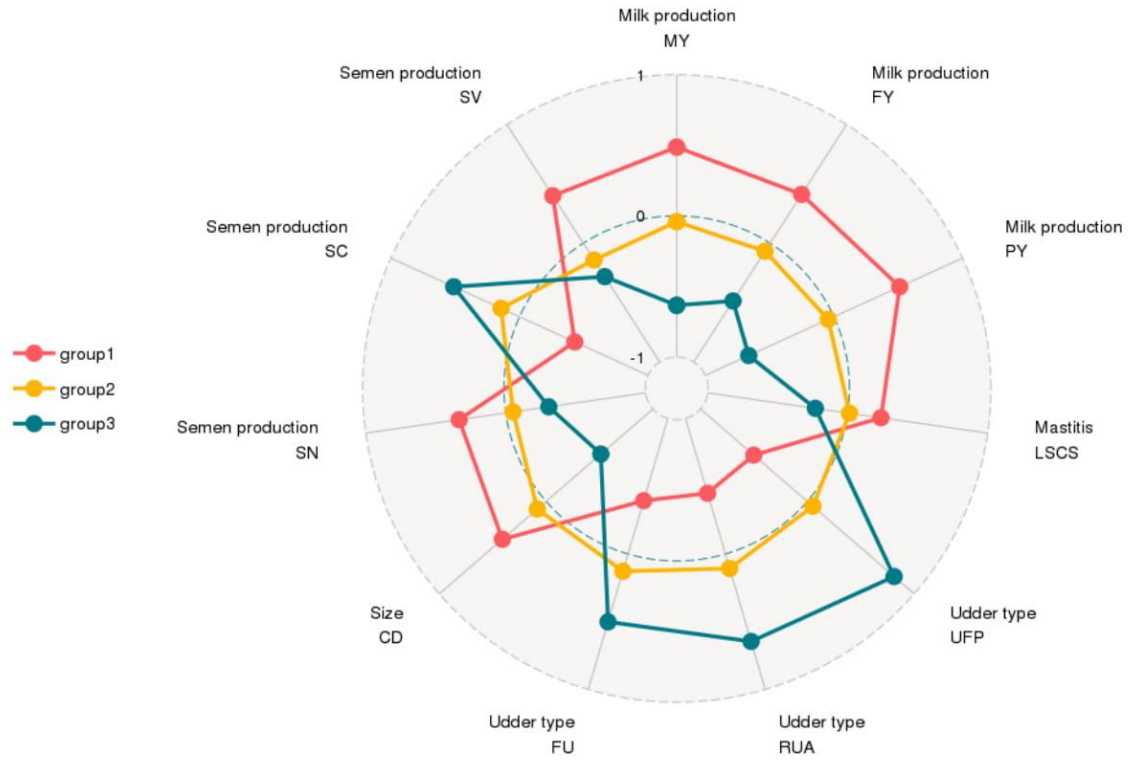


Figure 5. Radian plot showing the different daughter yield deviation (DYD) profiles of French Saanen based on the group attribution

DYD were standardized and normalized prior to the analysis

CD: chest depth, FU: fore udder, FY: fat yield, LSCS: lactation average somatic cell score, MY: milk yield, PY: protein yield, RUA: rear udder attachment, SC: semen concentration, SN: number of spermatozoa, SV: semen volume, UFP: udder floor position



Tables

Table 1 Summary of significant variants identified using imputed sequence data

534 **Table 2. Mean daughter yield deviation (DYD) per group of Saanen for the 33 Saanen animals with whole-genome sequence data**

535

	Number of individuals	MY	FY	PY	LSCS	UFP	RUA	FU	CD	SN	SC	SV
Group 1	10	99.83	3.43	3.62	0.14	-0.16[a]	-0.21	-0.19[a]	0.46	-0.02	-0.23[a]	1.04
Group 2	16	106.83	3.17	3.50	0.06	-0.01[a]	-0.03	-0.05[b]	-0.01	0.12	0.00[a]	0.75
Group 3	5	66.29	2.38	2.31	-0.06	0.39[b]	0.42	0.26[b]	0.09	-0.20	0.31[b]	0.62

536 a and b represent significant differences between groups ($P < 0.05$)

537 CD: chest depth, FU: fore udder, FY: fat yield, LSCS: lactation average somatic cell score, MY: milk yield, PY: protein yield, RUA: rear

538 udder attachment, SC: semen concentration, SN: number of spermatozoa, SV: semen volume, UFP: udder floor position

539

Table 3. Genotype for each of the three markers selected on the Illumina GoatSNP50 BeadChip for group attribution

Table 4. Specificity and sensibility of the Saanen profile attribution using three Illumina GoatSNP50 BeadChip SNPs for group attribution

Table 5. Results of profile assignment using three SNPs from the Illumina GoatSNP50 BeadChip.

Additional Files

Additional File 1 Table S1

Format: Excel data sheet

Title: List of genes associated to the QTL region of CHI19 in the Saanen breed

Description: List of genes found in the annotation of variants significantly associated to CHI19

Additional File 2 Table S2

Format: Excel data sheet

Title: Mean and standard deviation of each studied trait in the Saanen breed

Description: Mean and SD for the traits associated with chromosome 19

CD: chest depth, FU: fore udder, FY: fat yield, LSCS: lactation average somatic cell score, MY: milk yield, PY: protein yield, RUA: rear udder attachment, SC: semen concentration, SN: number of spermatozoa, SV: semen volume, UFP: udder floor position

II. Analyses et résultats complémentaires

II.1. Exploration d'un variant structural

L'exploration de fichiers *bam* disponibles en Alpines et Saanen françaises nous a permis d'identifier une délétion située entre les positions 23 677 412 et 23 678 780 sur le chromosome 19. La Figure 24 présente une capture d'écran de l'alignement des lectures pour 6 Saanen : 5 du groupe 1 et 1 du groupe 3 (cf groupes désignés dans l'article précédent) et 2 Alpines. On note qu'aucune lecture ne s'aligne dans cette région chez les 2 Alpines et la Saanen du groupe 3. Cet indel se trouve en amont de la région QTL et est proche d'un gène LOC102182725 localisé entre 23 675 883 et 23 676 835. Ce gène correspond à un récepteur olfactif notamment impliqué dans la chimiotaxie des spermatozoïdes (source : www.genecards.org). Ceci nous a conduit à lancer le génotypage des animaux pour lesquels de l'ADN étaient encore disponible. Ainsi, nous avons effectué des PCR sur les 11 Alpines et 9 Saanen séquencées ainsi que sur 1 150 autres Saanen génotypées sur puce 50K (163 mâles et 987 femelles). A ces génotypages s'est ajouté le *calling* des variants structuraux effectué par Thomas Faraut (INRAE, GenPhySE) sur l'ensemble des séquences françaises. Aux effectifs précédents s'ajoutent donc 45 Alpines et 24 Saanen. Nous nous sommes ensuite concentrés sur les Saanen. Après vérification de la qualité d'imputation, nous avons imputé le génotype pour l'indel pour l'ensemble des animaux génotypés avec la puce 50k. Ces génotypes imputés nous ont permis d'effectuer des analyses d'association. Plusieurs analyses ont été effectuées à l'aide du logiciel GCTA : (1) dans un premier temps, des analyses à l'aide d'un modèle polygénique classique (2) enfin des analyses qui incluent le variant structural en tant qu'effet fixe dans le modèle. Le variant structural n'a été significatif pour aucun des caractères étudiés. De plus, son ajout en tant qu'effet fixe n'a pas effacé les signaux observés précédemment dans la région du QTL. Nous en avons donc conclu que ce dernier n'est pas le variant responsable du QTL.

Figure 24: Visualisation sous IGV (Robinson et al., 2011) des lectures de 5 Saanen et 3 Alpines dans la région d'un variant structural

II.2. Mise à jour de la puce 50k et utilisation des nouveaux génotypes

II.2.a. Choix des marqueurs additionnels sur le chromosome 19

Suite à de nombreux échanges entre Illumina et l'IGGC, une mise à jour de la puce caprine actuelle a été envisagée. Un add-on d'environ 5 000 marqueurs a été développé. Pendant ma thèse, j'ai pu participer à la mise à jour de cette puce en sélectionnant des marqueurs dans les régions QTL et en particulier dans celle du chromosome 19. La sélection des SNP s'est faite suite à des analyses d'association préliminaires effectuées au début de l'année 2019. Le signal a été subdivisé en 7 zones qui correspondent à des pics différents au sein du signal (Figure 25). Les nouveaux SNPs ont été choisis dans chacune de ces zones. La sélection s'est faite selon les critères suivants: (1) significativité dans les analyses (p-value), (2) nombre de caractères pour lequel le variant est significatif, (3) profil de MAF en Saanen française, (4) couverture du signal, et enfin (5) proximité avec les SNP déjà présents sur la puce et significatifs dans les analyses. Plusieurs allers-retours ont été nécessaire avec l'outil en ligne d'Illumina pour estimer la qualité des sondes et remplacer les marqueurs dont les sondes étaient de trop piètre qualité. Finalement ce sont 177 SNP qui ont été ajoutés sur l'add-on de la puce par cette méthode. A ces derniers s'ajoute un variant candidat identifié par Stéphane Fabre (INRAE, GenPhySE) pour la fertilité de la semence dans le gène SHBG en position 27 087 979.

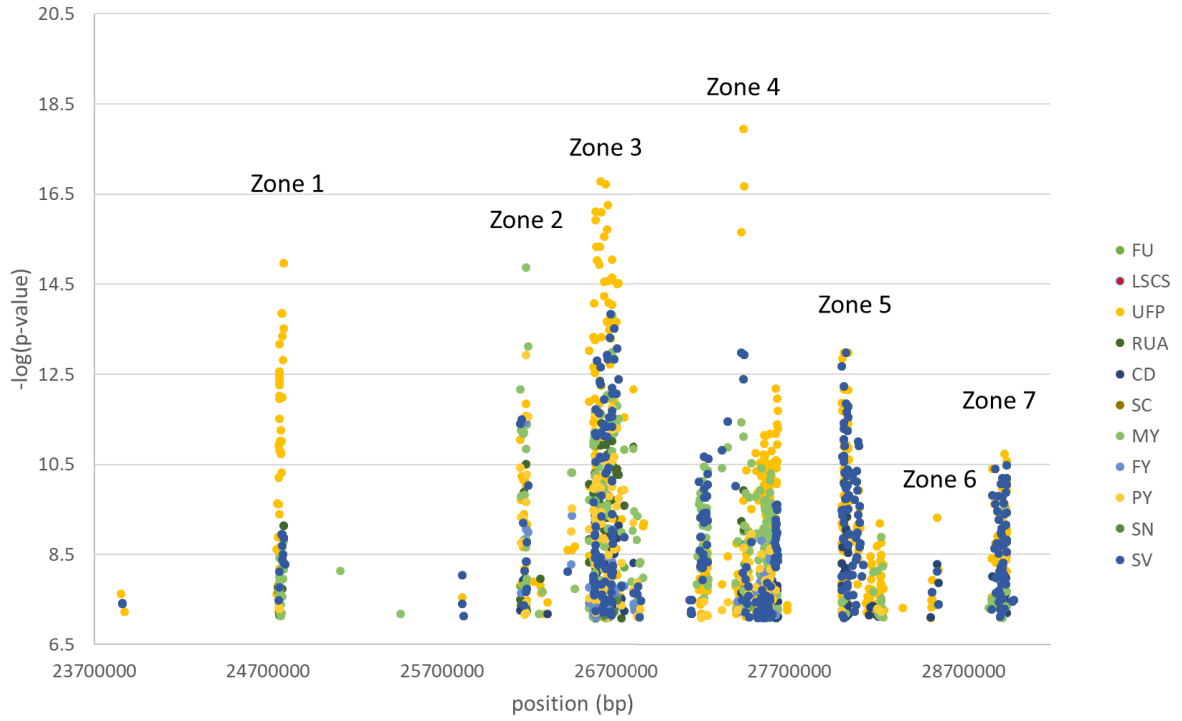


Figure 25: Découpage du signal dans la région du QTL en 7 zones

abréviations : CD : tour de poitrine ; FU : avant-pis ; FY : matière grasse ; LSCS : cellules somatiques du lait ; MY : lait ; PY : matière protéique ; RUA : qualité de l'attache-arrière ; SC : concentration en spermatozoïdes ; SN : nombre de spermatozoïdes ; SV : volume de semence ; UFP : position du plancher

II.2.b. Analyses d'associations sur la puce 50kv2

Parmi les individus génotypés pour valider les SNP de l'addon de la puce, on compte 53 Saanen. Ces individus ont pu être génotypés pour 168 des 177 marqueurs que nous avions sélectionnés. Nous avons profité de cette opportunité pour imputer ces SNP à l'ensemble des génotypes Saanen disponibles sur la première version de la puce. En premier lieu, un filtrage basique a été appliqué aux 168 SNP de l'addon correspondant à notre sélection. Ainsi, tous les SNP qui avaient un *call rate* inférieur à 95% ou une MAF inférieure à 1% ont été retirés du jeu de données. L'analyse s'est donc poursuivie sur 165 SNP. L'imputation est effectuée à l'aide de FImpute et du pedigree. Un scenario de *leave-one-out* a été mis en place pour tester la qualité de cette imputation. Le taux de concordance entre génotype vrai et génotype imputé a donc été mesuré tour à tour pour les 53 Saanen. Le taux de concordance a été de 100% pour tous les animaux à l'exception d'un individu (FRCH-SAA-0040). Ce dernier s'est avéré être en réalité une Alpine (erreur d'attribution d'identifiant lors de l'envoi des échantillons) et a été écarté du reste des analyses. Une fois cette vérification réalisée, nous avons imputé 1 574 Saanen qui étaient génotypées en v1 uniquement ou en v2 uniquement.

Enfin, nous avons effectué des analyses d'association à l'aide du logiciel GCTA sur les mâles en utilisant des DYD pour toutes les performances mesurées sur les femelles et sur les YD pour les caractères de production de semence. Ces résultats se sont avérés peu concluants. En effet, les SNP de la puce 50kv1 sont plus significatifs que ceux de l'addon pour l'ensemble des caractères. A titre d'exemple, la Figure 26 présente un Manhattan plot pour la position du plancher.

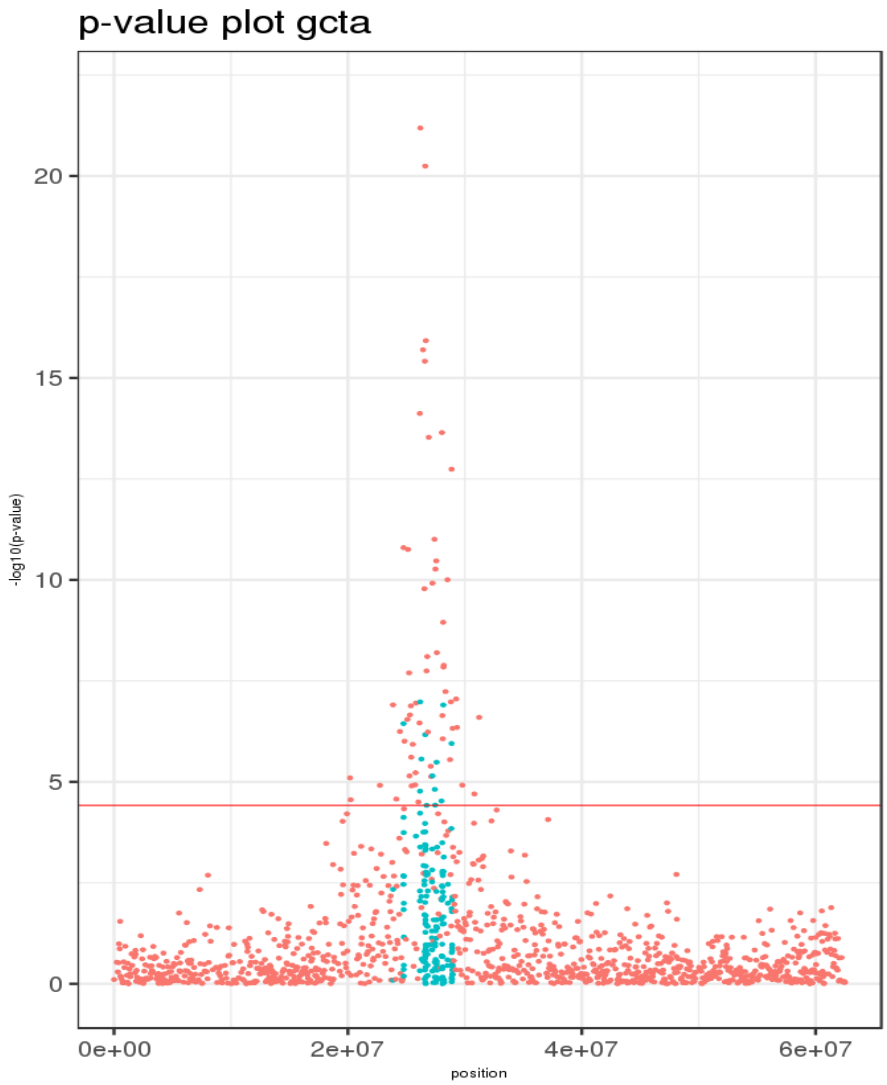


Figure 26: Manhattan plot pour le caractère de position du plancher (analyse effectuée sur des mâles ; en bleu : les SNP de l'addon ; en rouge les marqueurs de la puce v1)

Les SNP ayant été sélectionnés suite à des analyses d'association sur les performances propres des femelles génotypées, nous avons renouvelé l'analyse en utilisant les mêmes fichiers de performance. Le Manhattan plot pour le caractère position du plancher est présenté en Figure 27. On note que la tendance reste la même : les SNP de l'addon ne semblent pas

donner de meilleurs résultats que ceux de la première version de la puce. Notre sélection s'est faite sur des génotypes imputés de la puce 50kv1 vers la séquence, ce qui a probablement donné lieu à des erreurs dans les analyses d'association qui ont suivi. Ceci pourrait expliquer pourquoi ces SNP une fois génotypés ne sont pas à la hauteur de nos espérances.

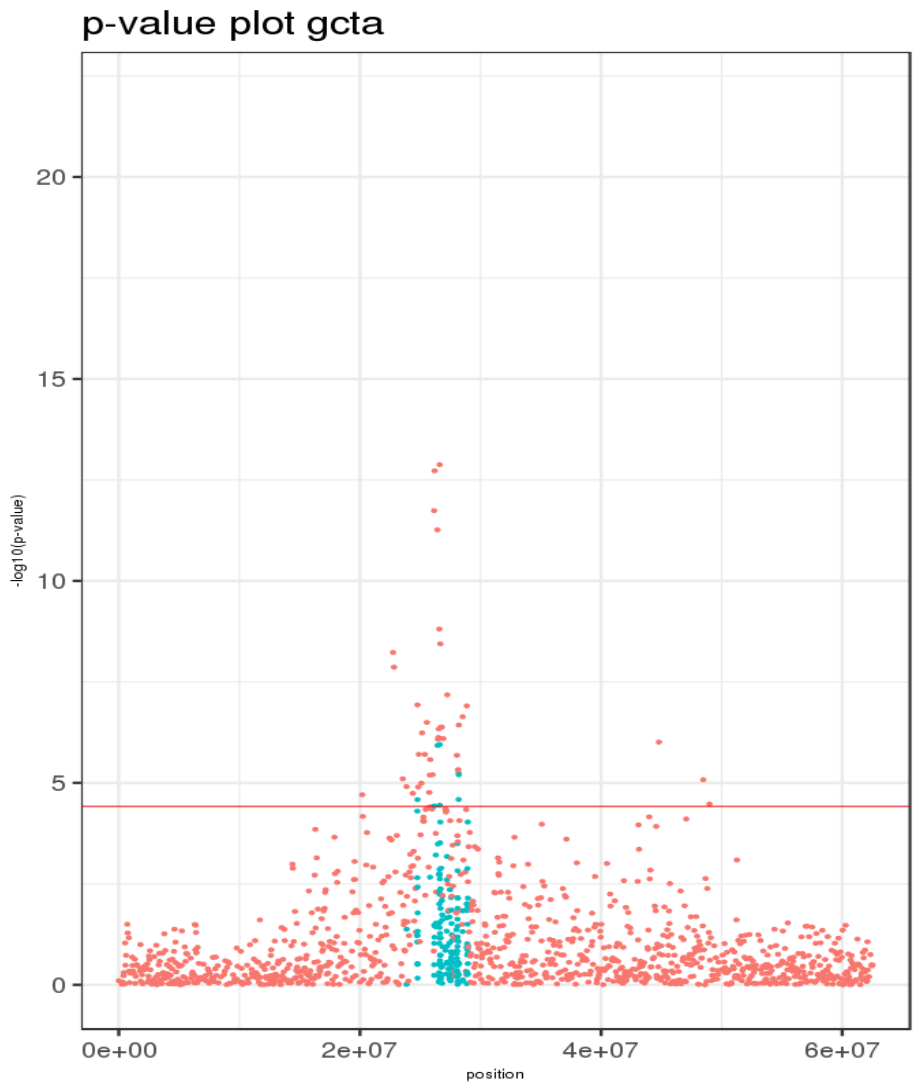


Figure 27: Manhattan plot pour le caractère de position du plancher (analyse effectuée sur des femelles ; en bleu : les SNP de l'addon ; en rouge les marqueurs de la puce v1)

II.3. Etude du déséquilibre de liaison dans la région du QTL

La densification de la région du QTL nous permet également de mieux estimer le déséquilibre de liaison de la région. Nous avons donc fusionné les génotypes 50kv1 et 50kv2 pour les 53 Saanen disponibles (à l'exclusion de FRCH-SAA-0040) puis extrait tous les marqueurs compris entre 23 et 30Mb sur le chromosome 19. Ces données ont été fournies à HaploView (Barrett, Fry, Maller, & Daly, 2005) qui a estimé la corrélation (R^2) pour

l'ensemble des paires de SNP. Les résultats sont présentés Figure 28 avec un gradient de gris, plus la case est foncée, plus le DL est fort. Sans surprise, le DL est fort lorsque les marqueurs sont proches. Les zones densifiées correspondent aux signaux QTL dans lesquels nous avons choisi des marqueurs. Ces zones sont en fort DL. Nous notons également des liaisons non-négligeables entre les différents pics. Ainsi en bas de la pyramide, le R n'est pas nul et peut même atteindre 1 pour certains marqueurs pourtant éloignés (visibles en bas à gauche sur la figure).

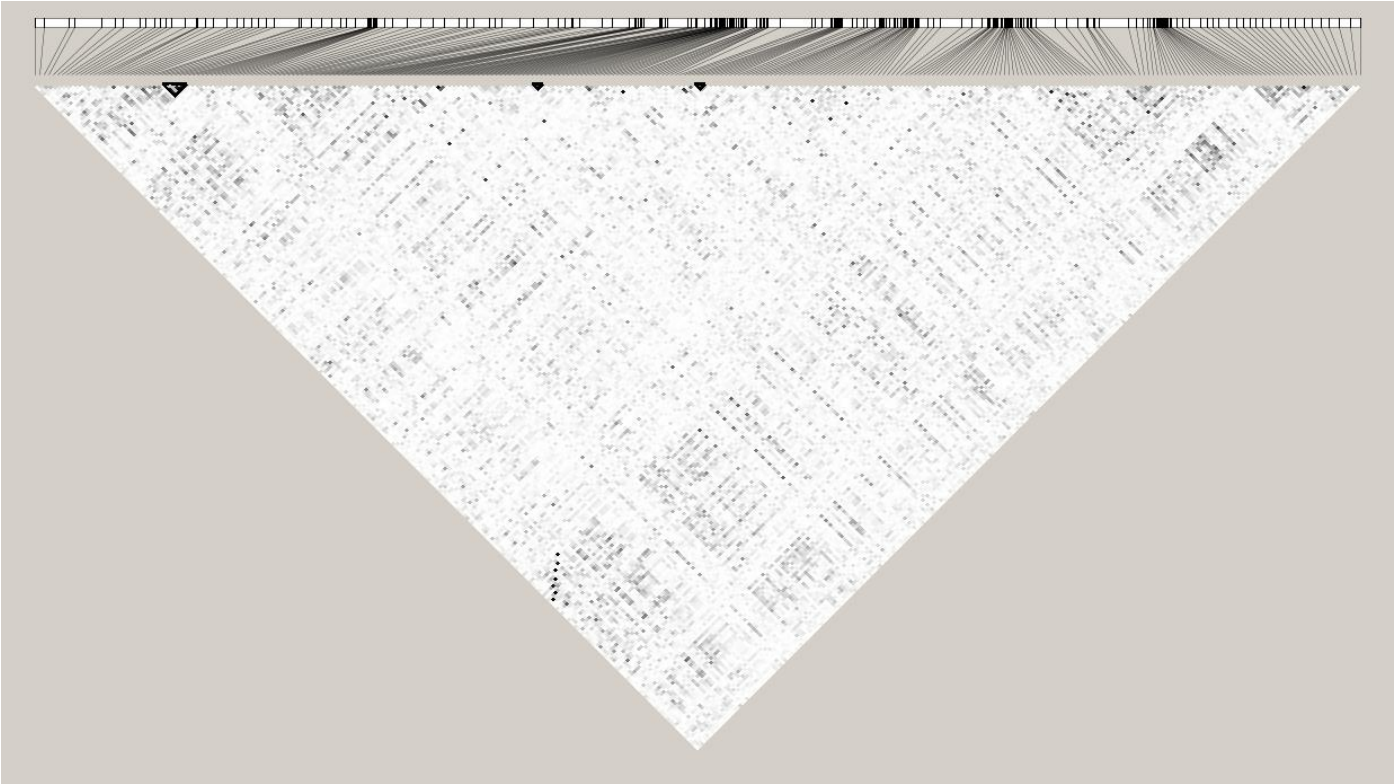


Figure 28: Bloc LD sur la région du QTL (entre 23 et 30 Mb) à partir des génotypes 50kv1 et 50kv2 (plus la case est foncée, plus le DL est fort)

Pour confirmer ces résultats, nous avons estimé le DL moyen sur le chromosome 19 et l'avons comparé avec le DL sur l'ensemble du génome. Cette analyse a été effectuée sur les marqueurs de la puce 50kv1. Le DL a été calculé en comparant un SNP avec ces 500 voisins (d'un côté comme de l'autre). Le DL de SNPs situés à des distances équivalentes a ensuite été moyenné par fenêtre de 50kb. La Figure 29 présente les résultats de l'analyse effectuée sur l'ensemble des génotypes 50k disponibles en Saanen (1 570). Ainsi, le DL décroît moins vite

avec la distance sur le chromosome 19 que sur le reste du génome et atteint un plateau légèrement plus élevé.

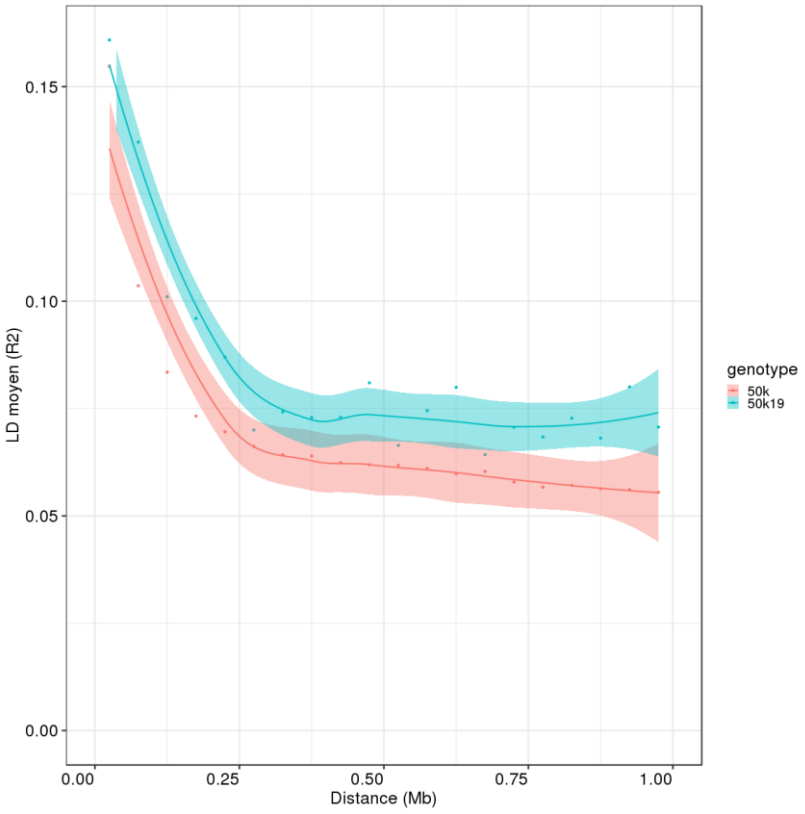


Figure 29: DL moyen entre 2 marqueurs de la puce 50kv1 en fonction de leur éloignement Génome entier en rouge, CHI19 en bleu

III. Conclusion du chapitre

Dans ce chapitre, notre objectif était de mieux comprendre la région de 5Mb du chromosome 19 associée à plusieurs caractères en sélection dans la race Saanen. Nous avons cherché à étudier le plus complétement possible cette région à partir des données génomiques disponibles : séquences brutes, séquences imputés et génotypes 50k.

Des analyses ACP sur séquence nous ont permis d'identifier des profils génétiques qui correspondent à différents profils en performance. Ces derniers semblent cohérents avec ce qui a été observé jusqu'à présent : corrélations négatives entre caractères de production et de conformation, plus faible progrès génétique pour les caractères de morphologie mammaire. Il est possible à partir d'une combinaison de marqueurs de la puce 50k d'attribuer un profil à un individu. Ceci pourrait permettre aux sélectionneurs de mieux définir leurs priorités (entre caractères de production et morphologie de la mamelle) et ainsi choisir les animaux qui leur

convienne le mieux pour continuer à améliorer les caractères de production sans dégrader les caractères de morphologie.

De nombreuses méthodes pour affiner la région QTL et proposer des variants et/ou des gènes candidats ont été testées. Ainsi nous avons effectuée des analyses d'association selon un modèle polygénique classique ou selon un modèle bslmm (Bayesian Sparse Linear Mixed Model). En parallèle, nous avons implémenté des analyses d'association sur des états de phases locaux en utilisant la suite de logiciel LinkPhase et ASREML (Gilmour, Thompson, & Cullis, 1995). Ces méthodes ne nous ont pas permis d'améliorer significativement la localisation du QTL ni d'identifier des gènes candidats.

Enfin, un balayage manuel des fichiers *bam* disponibles sur la région du QTL nous a permis de mettre en évidence un variant structural de grande taille en amont de la région QTL (entre 23 677 412 et 23 678 780). Ce dernier a fait l'objet d'un *calling* sur les données de séquence et d'un génotypage pour l'ADN encore disponibles pour les Saanen et quelques Alpines. Les analyses d'association sur ce variant ont rejeté l'hypothèse d'une association causale avec les caractères d'intérêt.

La Saanen française semble se différencier des autres races laitières de métropole sur cette région particulière du génome. A l'échelle de l'Europe, les Saanen françaises et italiennes semblent différer des Saanen suisses alors que ces dernières ont une origine commune. Enfin au niveau mondial, la Saanen française et la race mixte néo-zélandaise semblent porter le même QTL sur le chromosome 19. En revanche, la Saanen canadienne se rapproche plus de ce que nous avons pu observer en Alpine. La région présente un fort déséquilibre de liaison. Nous avons été en mesure de proposer une méthode pour identifier des profils différents de Saanen à partir d'un génotype 50k. Cette méthode pourrait servir aux organismes de sélection français comme internationaux. Ils seraient alors en mesure de sélectionner les individus qui présentent le profil le plus cohérent avec leurs objectifs de sélection et ce avant leur entrée en testage.

Chapitre 4
Etude de l'intégration des informations révélées par la séquence dans les évaluations génomiques en race Saanen

Dans cette partie, nous utiliserons les données de séquences et les résultats des études présentées précédemment dans les évaluations génomiques en race Saanen. Cette race est en effet porteuse d'une vaste région QTL sur le chromosome 19 associée à plusieurs caractères en sélection. Cette région nous permet donc d'étudier l'intérêt d'intégrer des données de séquence dans les évaluations tout en limitant les temps de calculs. Plusieurs stratégies ont été envisagées en lien avec le QTL détecté sur le chromosome 19 : ajout des variants de l'ensemble du chromosome, ajout des variants de la région QTL uniquement, ajout de la sélection de variants que nous avons proposée pour l'add-on de la puce caprine (Chapitre 3), ajout d'une sélection aléatoire d'un sous-jeu de variants. Toutes ces stratégies ont été comparées à un modèle single step (ssGBLUP) sur génotype 50k (modèle actuellement utilisé en routine). Enfin dans la lignée des travaux de thèse de Marc Teissier (Teissier, 2019), les modèles utilisant des pondérations de variants dans les modèles single step (WssGBLUP et WssGBLUP) ont été testés sur les différents sous-jeux de données.

Cette étude a fait l'objet d'un article scientifique accepté dans le journal *Journal of Dairy Science* le 11/08/2020. Quelques analyses supplémentaires viennent compléter ce chapitre : intégration de l'information « groupe » identifiée au chapitre 3 (3 profils de Saanen distinguables à partir de 3 SNPs) dans les évaluations génomiques et étude des biais des modèles génomiques ssGBLUP et WssGBLUP sur les jeux de données testés.

I. Effet de l'inclusion de données de séquence dans les évaluations génomiques – Article

I.1. Introduction et résumé de l'article

Les données de séquence représentent une perspective importante pour les évaluations génomiques. En effet, elles permettraient en théorie de s'affranchir du déséquilibre de liaison qui existe entre la mutation causale responsable des variations observées sur un ou plusieurs caractères d'intérêt et le marqueur le plus proche sur l'outil génomique utilisé dans les

évaluations (cf Chapitre 1). De ce fait, les évaluations génomiques sur données de séquences représentent une piste intéressante car la persistance de la qualité des évaluations au fil des générations pourrait être plus élevée (B. J. Hayes et al., 2014). L'exhaustivité de ces données est également intéressante car il est alors possible d'inclure dans les évaluations des types de variants qui sont traditionnellement écartés lors de la conception des puces, tels que insertions/délétions, petites MAF, etc… (Albrechtsen, Nielsen, & Nielsen, 2010), alors qu'ils peuvent avoir un impact non-négligeable d'un point de vue génétique (cf Chapitre 1).

L'addition de variants de séquence dans les évaluations génomiques a été étudié dans d'autres filières animales. En bovins laitiers (B. J. Hayes et al., 2014) comme en ovins (Moghaddar et al., 2018), les gains de précision ont été marginaux : entre 1,4 et 2,6% par rapport à une évaluation utilisant les génotypes 50k.

En caprins, les travaux de thèse de Céline Carillier ont évalué la faisabilité des évaluations génomiques en races Alpine et Saanen (Carillier et al., 2017) et ont permis dès 2018 leur mise en œuvre en routine via un modèle ssGBLUP. Les perspectives actuelles reposent principalement sur l'intégration de l'information de l'architecture génétique des caractères d'intérêts pour la filière. Partant du modèle instauré en routine (ssGBLUP), les travaux de thèse de Marc Teissier (Teissier, 2019) ont exploré des modèles alternatifs : WssGBLUP et ses variantes, et *gene content* notamment. Ces types de modèle sont intéressants en race Saanen car des QTL ont été identifiés pour 6 des 11 caractères évalués (Grosclaude et al., 1987; Martin et al., 2017). L'ensemble de ces travaux s'est appuyé sur les génotypes disponibles : génotypes 50k et génotypages de la région des caséines (CHI6) réalisé par le laboratoire Labogena à l'aide d'enzymes de restriction. Inclure le génotypage pour la caséine alphaS1 (gène CSN1S1) a conduit à des améliorations de la précision des évaluations comprises entre 6 et 27% en fonction du caractère (Carillier-Jacquin et al., 2016). Dans le cas du WssGBLUP, les caractères les plus impactés sont ceux pour lesquels un QTL a précédemment été identifié : TP (CHI6) (Teissier et al., 2018), TB (CHI14), avant-pis (CHI19), position du plancher (CHI19), qualité de l'attache arrière (CHI19), LSCS (CHI19), lait (CHI19), MG (CHI19), MP (CHI19) (Teissier et al., 2019).

A l'heure actuelle, aucune étude portant sur des évaluations génomiques utilisant des données de séquence n'a encore été publiée en caprins. Nos travaux n'ont pas permis d'identifier de mutation causale, l'approche *gene content* ne nous a donc pas paru être appropriée. Nous nous sommes alors concentrés sur les alternatives pondérées au ssGBLUP. Nous n'avons pas identifié de région QTL majeure en race Alpine et les QTL précédemment

identifiés dans cette race dans le cadre des travaux de thèse de Pauline Martin (Martin et al., 2018; Martin et al., 2017) ont été largement étudiés dans le cadre de la thèse de Marc Teissier. Nous avons donc restreint notre étude à la région QTL du chromosome 19 identifié en Saanen, région qui a, par ailleurs, fait l'objet d'une cartographie fine dans cette thèse (Chapitre 3).

Dans notre étude, plusieurs scenarios ont été testés ajoutant chacun un sous-ensemble de variants du chromosome 19 aux génotypes 50k. Ainsi, 5 ensembles de variants ont été utilisés : (1) aucun variant additionnel (scenario de référence) (2) ajout de 178 variants sélectionnés entre 23 à 30 Mb sur le chromosome 19 pour l'add-on de la puce v2 (3) ajout de 178 variants sélectionnés au hasard sur le chromosome 19 (4) ajout de l'ensemble des 586 623 variants imputés sur le chromosome 19 (5) ajout de l'ensemble des 69 416 variants de la région QTL du chromosome 19. Et 3 modèles ont été envisagés : le ssGBLUP, le WssGBLUP et le WssGBLUP utilisant des fenêtres de 40 variants consécutifs ou des fenêtres de 2,4 Mb. Les caractères évalués étaient les caractères de morphologie mammaire : avant-pis, orientation des trayons, position du plancher, profil de la mamelle, qualité de l'attache arrière, la santé de la mamelle avec les comptages de cellules somatiques du lait ainsi que 3 caractères de productions associés au chromosome 19 : le lait et les matières protéique et butyreuse.

Les meilleurs résultats ont été obtenus en utilisant un ssGBLUP incluant les génotypages 50k et les variants imputés de la région du QTL du chromosome 19 restreinte aux variants compris entre 24,72 et 28,38 Mb : +6,2% de précision en moyenne sur les caractères évalués. Le gain de précision le plus important a été obtenu pour la matière grasse (+ 17,9%). Pour le lait et la matière protéique, les améliorations étaient de +13.7% et 12,5% respectivement. Enfin les évaluations des caractères de morphologie associés à la région du QTL se sont avérées plus précises avec l'intégration des variants de la région. Malgré son association avec la région du QTL, les prédictions pour les cellules se sont trouvées légèrement dégradées par l'ajout de variants imputés (-7,1%). L'ajout de l'ensemble des variants du chromosome 19 a conduit à une réduction significative de la précision des prédictions : -4.8% en moyenne. Enfin la mise-à-jour de la puce caprine nous paraît être prometteuse car elle permet d'améliorer significativement la qualité des évaluations génomiques (entre 3,1 et 6,4% en fonction du scenario considéré) tout en limitant les temps de calculs liés à l'imputation notamment. Elle permettrait, de plus, d'obtenir des génotypes fiables.

I.2. L’utilisation de variants issus de la séquence d’une région QTL améliore la précision des évaluations génomiques en chèvres Saanen françaises : Article

Interpretive summary

Using sequence variants of a QTL region improves the accuracy of genomic evaluation in French Saanen goats

Talouarn E.

The recent decrease in sequencing costs has made it possible to sequence large numbers of individuals in key livestock species. Sequence data is interesting for genomic evaluations as it can include causal mutations and therefore lead to a better persistency of the accuracy of genomic predictions. Here, we describe the first exploratory study of genomic evaluations using information extracted from sequence data in French Saanen goats. Our study shows that selected sequence data significantly improves the accuracy of genomic predictions in the French Saanen breed.

Running head: GENOMIC EVALUATION OF FRENCH DAIRY GOATS

Using sequence variants of a QTL region improves the accuracy of genomic evaluation in French Saanen goats

Estelle Talouarn[*,1], Marc Teissier[*], Philippe Bardou[†], Hélène Larroque[*], Virginie Clément[‡], Isabelle Palhière[*], Gwenola Tosser-Klopp[*], Rachel Rupp[*] and Christèle Robert-Granié[*]

[*]GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet-Tolosan, France

[†] Sigenae, INRAE, F-31326 Castanet-Tolosan, France

[‡] Institut de l'Elevage, F-31326 Castanet-Tolosan, France

[1]Corresponding author: Estelle Talouarn

Mailing address: GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet-Tolosan, France

E-mail: estelle.talouarn@inrae.fr

ABSTRACT

The enhanced availability of sequence data in livestock provides an opportunity for more accurate predictions in routine genomic evaluations. Such evaluations would therefore no longer rely only on the linkage disequilibrium between a chip marker and the causal mutation. The objective of this study was to assess the usefulness of sequence data in Saanen goats (N = 33) to better capture a QTL on chromosome 19 (CHI19) and improve the accuracy of predictions for three milk production traits, five type traits and somatic cell scores. All 1,207 50k genotypes were imputed to the sequence level. Four scenarios, each using a subset of CHI19 imputed variants, were then tested. Sequence derived information included all CHI19 variants (529,576), all variants in the QTL region (22,269), 178 variants selected in the QTL

region and added to an updated chip, or 178 randomly selected variants on CHI19. Two genomic evaluation models were applied: single-step GBLUP and Weighted single-step GBLUP. All scenarios were compared with ssGBLUP using 50k genotypes.

Best overall results were obtained using single-step GBLUP on 50k genotypes completed with all variants in the QTL region of chromosome 19 (+6.2% average increase in accuracy for nine traits) with the highest accuracy gain for fat yield (+17.9%), significant increase for milk (+13.7%) and protein yields (+12.5%) and type traits associated with CHI19. Despite its association with the QTL region of chromosome 19, the somatic cell score showed decreased accuracy in every alternative scenario. Using all CHI19 variants led to an overall decrease of 4.8% in prediction accuracy. The updated chip was efficient and improved genomic evaluations by 3.1% to 6.4% on average depending on the scenario. Indeed, information from only a few carefully selected variants increased accuracies for traits of interest when used in a ssGBLUP model.

In conclusion, using QTL-region variants imputed from sequence data in single-step genomic evaluations represents a promising perspective for such evaluations in dairy goats. Furthermore, using only a limited number of selected variants in QTL regions, as available on SNP chip updates, significantly increases the accuracy for QTL-associated traits without deteriorating the evaluation accuracy for other traits. The latter approach is interesting, as it avoids time-consuming imputation and data formatting processes while providing reliable genotypes.

Keywords: genomic evaluations; sequence; Saanen; dairy goats

INTRODUCTION

The recent decrease in sequencing costs has made it possible to sequence large numbers of individuals in livestock species. Including sequence data in genomic evaluations is interesting because it might improve the persistency of genomic prediction accuracy (Hayes et al., 2014).

Indeed, such evaluations would no longer rely simply on the linkage disequilibrium between a chip marker and the causal mutation of a trait. In sheep (Moghaddar et al., 2018) and dairy cattle (Hayes et al., 2014), the gain in accuracy when sequence data were included in the kinship matrix calculation ranged from 1,4% to 2,6% compared with genomic evaluations performed on 50k genotypes and reached up to 2% when compared with high-density (**HD**) genotypes in dairy cows.

The VarGoats program made available over 1,000 sequences of the *Capra* genus from 125 breeds (http://www.goatgenome.org/vargoats.html). Talouarn et al. (Talouarn et al., 2020) investigated the feasibility of imputation from 50k-chip information using the sequence data available for 33 sequenced French Saanen individuals. Acceptable imputation accuracy was achieved within-breed with mean allele and genotype concordance rates of 0.86 and 0.74 respectively in the Saanen breed. The imputation to sequence data also fine mapped a previously identified **QTL** in Saanen goats and led to the identification of new signals in both Alpine and Saanen breeds (Talouarn et al., 2020). The QTL region of chromosome 19 (CHI19) explains between 5 and 10% of the total additive genetic variance of somatic cell score (SCS) and type traits (Martin et al., 2018). The refined information for CHI19 provided by sequence data might be of interest when performing genomic evaluations in Saanen goats because 6 of the 11 traits included in the indexes are associated with the QTL (Martin et al., 2018; Martin et al., 2017; Talouarn et al., 2020).

France is a leading European country in goat milk production. Selection strategies aim at improving cheese-making. A synthetic production index has been established using production traits: milk yield, protein and fat yields, protein and fat contents. Type traits such as fore udder, teat orientation, udder floor position, udder profile and rear udder attachment are now also included in a morphology index. Indexes for both production and type traits are combined in a synthetic index which differs between Alpine and Saanen breeds (Clément et al., 2006;

85 Larroque et al., 2011). Since 2013, udder health is considered for AI buck selection using the
86 somatic cell count (Virginie Clément, Institut de l'Elevage, personal communications).
87 Following the introduction in 2011 of the 50k genotyping chip (Illumina GoatSNP50
88 BeadChip) (Tosser-Klopp et al., 2014), the feasibility of genomic evaluations in French dairy
89 goats has been studied (Carillier et al., 2013). Such evaluations were officially implemented in
90 French Alpine and Saanen in 2018 using a single-step **GBLUP** model (**ssGBLUP**) (Legarra et
91 al., 2009). Current perspectives for improving genomic evaluations rely on finding better ways
92 of integrating genotype information. Studies were therefore performed to include major gene
93 information in the evaluations using Weighted single-step GBLUP (**WssGBLUP**) (Teissier et
94 al., 2019; Teissier et al., 2018). SNPs close to the casein α-S1 gene had higher weights and led
95 to substantial gains in accuracy for the genomic estimated breeding values (**GEBV**) of protein
96 content in Alpine and Saanen goats (Teissier et al., 2018). As mentioned, other studies identified
97 a large QTL region on CHI19 for udder type, udder health and milk production traits (Martin
98 et al., 2018; Martin et al., 2017; Mucha et al., 2017) in Saanen and mixed-breed goats.
99 WssGBLUP led to gains in accuracy ranging from 2 to 14% in Saanen goats for traits associated
100 with the QTL region of CHI19 (Teissier et al., 2019).

101 Here we describe the first exploratory study of genomic evaluations using information extracted
102 from sequence data in French Saanen goats. Our objective was to take advantage of available
103 sequence data (N = 33) to include refined CHI19 information in French Saanen genomic
104 evaluations for nine traits: milk yield, fat and protein yields, fore udder, teat orientation, udder
105 floor position, udder profile, rear udder attachment and somatic cell score. We investigated the
106 relevance of including sequence variants in routine genomic evaluations for these nine traits.
107 We focused on identifying the method that would maximize both the accuracy of prediction
108 and the computation efficiency.

MATERIAL AND METHODS

This study did not require ethical approval because no experiments on animals were necessary (samples originated from other studies).

Animals, phenotypes and 50k genotypes

Details on phenotypes and 50k-genotype quality checks are described in Teissier *et al.* (Teissier et al., 2018). The milk performance dataset was provided by the French national milk records system and the udder traits records by the breeding company Capgenes. Phenotypes for milk production were considered over the whole lactation period: 250-d milk yield (**MY**, in kg), 250-d protein and fat yields (**PY** and **FY** respectively, both in kg), 250-d somatic cell score (**LSCS**). The type traits were: fore udder (**FU**), teat orientation (**TO**), udder floor position (**UFP**), udder profile (**UP**) and rear udder attachment (**RUA**). Type traits were only measured once per animal, mainly during their first lactation and sometimes in their second. Phenotypes, pedigree data, and genotypes were obtained from the official genetic evaluation of January 2016 (Larroque et al., 2011). Phenotype data was retained only for French Saanen goats born between 1980 and 2017. The final dataset is described for each trait in Table 1.

The pedigree consisted of 2,177,617 individuals and was complemented by defining unknown parent groups. Sixteen groups were defined according to the year of birth of the descendants: before 1975, between 1975 and 1980, between 1980 and 1983 and then every two years. Males and females were pooled together in unknown parent groups because there were few animals with unknown dams.

Genotypes were acquired with the Illumina GoatSNP50 BeadChip. A total of 1,207 genotypes (394 males, 813 females) were retained for French Saanen goats after the quality check step. The quality check step was previously described in Talouarn et al. (Talouarn et al., 2020). It implies removing all individuals with a call rate below 95% or showing pedigree inconsistency.

SNP quality control was based on the following inclusion criteria: call rate above 99%, Minor Allele Frequency (**MAF**) above 1% and Hardy-Weinberg p-value above 10^{-6}. After quality control, 47,147 SNPs were retained, including 1,143 SNPs on chromosome 19.

Sequence-derived information

Quality check of sequence data and imputation

Sequence data are described in Talouarn *et al.* (Talouarn et al., 2020). The sequence data were retrieved from the VarGoats project (http://www.goatgenome.org/vargoats.html). The 37 French Saanen individuals came from VarGoats child projects PRJEB37276, PRJEB37276 and NextGen PRJEB5900 projects. The final dataset comprised 33 French Saanen goats (31 males and 2 females), 4 individuals were removed as their mean coverage was below 5. All of these were also genotyped with the Illumina GoatSNP50 BeadChip.

A wide QTL region was previously identified in Saanen goats on CHI19 between 24.72 and 28.38 Mb (Martin et al., 2018; Mucha et al., 2017; Talouarn et al., 2020). This region is associated to production traits (milk yield, fat and protein yields), stature, udder type and health as well as semen volume. Sequence quality check and soft filtering processes were described in detail by Talouarn *et al.* (Talouarn et al., 2020) and resulted in keeping 23,337,436 variants including 539,476 variants on chromosome 19 and 22,269 variants between 24.72 and 28.38 Mb. Before imputing available 50k genotypes, imputation was necessary to fill in the gaps of the sequenced panel. Using a combination of AlphaImpute (v 1.9) (Hickey et al., 2012) and FImpute (v 3.0) (Sargolzaei et al., 2014) gave higher concordance rates between 50k genotypes and sequence than using solely one software while minimizing computation time. AlphaImpute (v 1.9) (Hickey et al., 2012)and FImpute (v 3.0) were therefore used. The mean concordance rate between 50k-genotypes and sequence data, calculated on the 33 sequenced Saanen goats, was 98.43% (±1.35) after filtering and imputation. No missing genotypes remained for sequenced Saanen after these steps.

158 Finally, imputation of the 1,207 50k-genotypes was performed using pedigree information on
159 chromosome 19. Mean genotype and allele concordance rates were estimated at 71.8% and
160 84.3% respectively in the Saanen breed.

161 *Illumina GoatSNP50 BeadChip update*

162 The Illumina GoatSNP50 BeadChip is currently being updated with a set of about 6,832 probes
163 (including duplicates) to be validated, resulting in the Goat IGGC 65k v2 chip. The region
164 between 23 and 30 Mb on chromosome 19 was densified to better capture the previously
165 identified QTL. The 178 variants in the QTL region added to the chip were chosen based on
166 association analysis using sequence information (E. Talouarn, unpublished results). Variants
167 were selected using the following criteria: (1) p-value in the association analysis, (2) number
168 of traits significantly linked to the variant, (3) MAF profile in French Saanen goats, (4) spacing
169 within the signal, and (5) distance to SNPs of the current version of the chip. The 178 positions
170 were extracted from imputed sequence data to quantify the impact of the increased
171 representation of chromosome 19 (+178 variants) on the accuracy of genomic evaluations. In
172 order to determine whether gains in precision were only due to the increased SNP density,
173 another 178 variants were also randomly selected on chromosome 19 using PLINK software
174 (Purcell et al., 2007). Among them, only 11 were located in the QTL region of chromosome 19.

175 ## *Evaluation methods*

176 ### *Single-step GBLUP (ssGBLUP).*

177 For single-step GBLUP, the model consisted in the following:

178 $$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + W\boldsymbol{p} + \boldsymbol{e} \quad \textbf{[1a]}$$

179 Where y is the vector of performances for the studied trait, β is the vector of fixed effects
180 defined as in the official genomic evaluations. For production traits and LSCS, fixed effects
181 were defined for each year and parity and were: herd, age at kidding x 4 geographic regions,

month of kidding x 4 regions, length of dry period x 4 regions. $\boldsymbol{u}$ is a vector of random additive genetic effects assumed to be normally distributed $N(0, H\sigma_u^2)$. $\boldsymbol{p}$ is the vector of random permanent environmental effects assumed to be normally distributed $N(0, I\sigma_p^2)$. $\boldsymbol{e}$ is a vector of random residuals also normally distributed $N(0, I\sigma_e^2)$. X is the incidence matrix relating phenotypes and the fixed effects. Z and W are the design matrices linking phenotypes to genetic and permanent environmental effects, respectively. The H matrix is the genetic relationship matrix which integrates both genotype and pedigree information implemented as in Legarra *et al.* (Legarra et al., 2009):

$$H = \begin{pmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{pmatrix}$$

Where A is a pedigree-based relationship matrix with indices 1 for ungenotyped animals and 2 for genotyped animals, and G is the genomic relationship matrix derived as in Christensen and Lund (Christensen & Lund, 2010):

$$G = 0.95\frac{M'M}{2\sum_{i=1}^{m} p_i(1 - p_i)} + 0.05A_{22}$$

Where m is the number of variants, p_i the estimated allele frequency at the locus i and M is a centered matrix of variant genotypes.

For udder traits, which were measured only once in the life of goats, the genomic evaluation model did not include a permanent environmental effect, and was as follows:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{e} \textbf{ [1b]}$$

Where $\boldsymbol{y}$, $\boldsymbol{u}$ and $\boldsymbol{e}$ are the same as previously stated. $\boldsymbol{\beta}$ includes 3 combined fixed effects: herd x parity x year, age at scoring x year and days in milk at scoring x year. Variance components were estimated by using the restricted maximum likelihood (**REML**) method in the remlf90

software (Misztal et al., 2002a) and ssGBLUP analyses were performed with the blup90iod2 software (Misztal et al., 2002b).

Weighted single-step GBLUP (WssGBLUP).

The same model was used to perform Weighted single-step GBLUP (WssGBLUP). WssGBLUP was applied to each trait studied using blupf90 family software (Misztal et al., 2002b). SNP effects and SNP weights were estimated using postGSf90 software. Our objective was to take advantage of the sequence information for chromosome 19 in order to increase the accuracy of genomic evaluations for the traits related to the QTL region. WssGBLUP is based on iterative ssGBLUP in which weights for SNP variances are used to form the genomic relationship matrix G. This method is used to give more weight to causal mutations, variants that are in high LD with a causal mutation or variants within a QTL region with a relatively large effect. The genomic relationship matrix G was built differently. The solutions for genomic breeding values from ssGBLUP can be decomposed into SNP effects as follows (Huiyu Wang et al., 2014):

$$\hat{a} = DM'(MDM')^{-1}\hat{u}_g$$

where $\hat{a}$ is a vector of variant effects, D is a diagonal matrix of weights (set to one in ssGBLUP), M is the centered matrix of variant genotypes and $\hat{u}_g$ the vector of GEBVs from genotyped animals only. The additive variances of the effect of variant i were estimated as:

$$\sigma^2_{u,i} = 2\hat{a}_i^2 p_i(1 - p_i)$$

Where p_i is the allele frequency of variant i. The vector of variances of SNP effects was normalized (the normalization process ensured that the sum of the variances remained constant and was equal to the number of SNPs) and used as weights in matrix D to construct the weighted matrix G (G^*) as described by Wang et al. (Huiyu Wang et al., 2014):

226 $$G^* = 0.95\frac{M'DM}{2\sum_{i=1}^{m} p_i(1-\ p_i)} + 0.05A_{22}$$

227 GEBVs were estimated again with Models [1a] and [1b] by considering weights for each SNP
228 via the G^* matrix included in the H matrix. This process was carried out iteratively with weights
229 estimated at each iteration as described by Wang et al. The following formulas were used as
230 described by Wang et al (Wang et al., 2012):

231 1. Iteration 1, initialization, $D_{(1)} = I$; $G^* = 0{,}95 * \lambda ZD_{(1)}Z' + 0{,}05 * A_{22}$
232 2. Calculation of G^* following the previous formula to obtain the EBV vector $\hat{u}_g$.
233 3. Iteration 2 (it)
234 4. Estimation of variant effects $\hat{a}_{(it)} = \lambda D_{(it-1)}Z'G^{*-1}_{(it-1)}\hat{u}_{g\ (it-1)}$
235 5. Conversion of effects into weights following the formula: $d_i^* = \hat{a}_i^2 * 2p_i(1-\ p_i)$.
236 Weights are integrated into matrix $D^*_{(it)}$.
237 6. Weights normalization $D_{(it)} = \frac{tr(D_{(1)})}{tr(D^*_{(it)})} * D^*_{(it)}$.
238 7. Building $G^*_{(it)}$. $G^*_{(it)} = 0{,}95 * \lambda ZD_{(it)}Z' + 0{,}05 * A_{22}$
239 8. Launch of a WssGBLUP using the newly calculated $G^*_{(it)}$ matrix to obtain new EBVs.
240 9. Exit

241 In French dairy goats, Teissier *et al.* (Teissier et al., 2018) showed that 2 iterations were
242 sufficient to maximize the accuracy. We therefore stopped after 2 iterations in our study. This
243 model will be referred to as standard WssGBLUP.

244 *Weighted single-step GBLUP using windows (WssGBLUP$_{windows}$).*

245 As proposed by Zhang et al. (Zhang et al., 2016), several other methods can be considered to
246 calculate the weight for variants in the D matrix. These methods assign the same weight to
247 several consecutive SNPs within a chromosomal region. These methods have already been

248 explored by Teissier *et al* (Teissier et al., 2018) in French dairy goats. The best results were
249 obtained when using non-overlapping windows of 40 consecutive SNPs and taking the
250 maximum weight in the window. These weights were calculated based on the weights estimated
251 with WssGBLUP. The highest weight observed in a window was assigned to all SNPs within
252 the same window. When adding sequence information for the QTL region on chromosome 19
253 (22,269 sequence variants between 23 and 30 Mb), the average spacing between variants was
254 much lower than on other chromosomes (2,670 bp ± 13,290 vs. 60,000 bp for the rest of the
255 genome). We decided to take this information into account when building the windows. We
256 tested two options: (1) using windows of 40 consecutive variants or (2) using windows of 2.4
257 Mb (average spacing of 40 consecutive SNPs on the chip).

258 *Tested scenarios*

259 Several scenarios were tested, based on either ssGBLUP or WssGBLUP methods, and using
260 four different sets of new markers in addition to the Illumina GoatSNP50 BeadChip (Table 2).
261 We wanted to see how we could expect the next version of the Illumina GoatSNP50 BeadChip
262 to perform if used for the genomic evaluations (geno_50kv2QTL). As mentioned previously,
263 the 178 additional variants selected in the QTL region of chromosome 19 and imputed in our
264 dataset were extracted for that purpose. In order to assess the relevance of the results, another
265 scenario consisted in including 178 randomly selected variants located over the whole length
266 of chromosome 19 (geno_50kv2_random). In scenarios 3 and 4, all imputed variants from the
267 sequence on chromosome 19 (N = 539,476) (geno_50kseqCHI19) or only those in the QTL
268 region (22,269) (geno_50kseqQTL) were chosen, respectively.

269 As our goal was to find the best candidate scenario to improve genomic predictions, all
270 scenarios were systematically compared with single-step genomic evaluation using 50k
271 genotypes (geno_50k).

Single-step GBLUP.

The different scenarios and information used are summarized in Table 2.

Weighted single-step GBLUP.

Our objective using WssGBLUP was to compare the integration of sequence data for the QTL region of chromosome 19, i.e. geno_50kseqQTL (22,269 sequence variants in the QTL region) or geno_50kv2QTL (178 SNPs in the QTL region) with a simple evaluation using 50k genotypes (geno_50k). This was performed using a standard WssGBLUP model.

We also attempted to use windows of consecutive markers ($WssGBLUP_{windows}$). The highest weight in the window was assigned to all the SNPs in the same window. Because the average spacing in the QTL region was smaller than when solely using 50k genotypes for all 4 approaches, we implemented windows of 2.4 Mb which is the average distance covering 40 SNPs with the Illumina GoatSNP50 BeadChip. This window size has proven to be efficient in previous studies using 50k genotypes (Teissier et al., 2018). We retained the alternative using windows of 40 consecutive SNPs for comparison purposes. We implemented 4 approaches: (1) WssGBLUP on 50k genotypes building windows of 2.4 Mb over the whole genome, (2) WssGBLUP on 50k genotypes and sequence information for the QTL region (either 178 selected variants or 22,269 sequence variants) building 2.4-Mb windows only on chromosome 19, (3) WssGBLUP on 50k genotypes and sequence information for the QTL region (either 178 selected variants or 22,269 sequence variants) building windows of 40 consecutive variants over the whole genome, and (4) the same scenario as (3) but building windows of 2.4 Mb.

Accuracy of genomic predictions.

The reference population used to assess the accuracy of genomic evaluation comprised only genotyped males, even if the genotypes of females were also used in ssGBLUP and WssGBLUP evaluations. This reference population was split into two subsets: a training set and a validation

296 set. The training population consisted of 248 sires born between 1998 and 2007 and genotyped
297 with the Illumina GoatSNP50 BeadChip. All the information for these animals (genotype,
298 pedigree with their ancestry and progeny, and phenotypes of their progeny) was retained in the
299 datasets to estimate the GEBVs. The validation population consisted of 146 bucks born between
300 2008 and 2012. All of their progeny with phenotypes were removed from the dataset. The
301 GEBVs estimated in these conditions and the **DYD**s (Daughter Yield Deviation) computed
302 from the official genetic evaluation of January 2016, were compared for the 146 animals in the
303 validation set. DYDs were the average performance values for the daughters corrected for fixed
304 and random environmental effects and half of the merit of their dams. DYDs were weighted by
305 effective daughter contributions as described by VanRaden and Wiggans (VanRaden &
306 Wiggans, 1991). The accuracy of genomic predictions was assessed by calculating the Pearson
307 correlation between the GEBVs estimated with each model and DYDs. To compare the Pearson
308 correlations obtained with the different scenarios and methods, we used the Hotelling-Williams
309 test as implemented in the *multilevel* R package (Williams, 1959).

310 # RESULTS

311 All tested scenarios were systematically compared with ssGBLUP using 50k genotypes
312 (geno_50k).

313 ## *ssGBLUP*

314 Figure 1 compares the accuracy of evaluations with the current version of the chip (geno_50K),
315 the updated version of the chip (178 additional variants, geno_50v2QTL), the 178 randomly
316 selected variants on chromosome 19 (geno_50kv2QTL_random) and evaluations using either
317 sequence data of the QTL region (geno_50kseqQTL) or the imputed variants of the whole
318 chromosome 19. The percentage of variance explained for each trait in the geno_50k,
319 geno_50kv2QTL and geno_50kseqQTL scenarios were calculated per variant in a ssGBLUP

model using the option *windows_variance* parameter of BLUPF90.Variance explained by the QTL region was obtained by summing the variance explained by each SNP and are shown in Figure 2. In the geno_50kseqQTL scenario, adding sequence information with a known chromosomal location led to significant gains in accuracy (+6.2% on average compared with geno_50k). The highest gain in accuracy was observed for FY (+17.9%) and significant increases were obtained for MY (+13.7%), PY (+12.5%), FU (+4.5%), UFP (9.0%), and RUA (4.9%). For LSCS, the accuracy decreased by 7.1% (Figure 1). However, when information was added over the whole chromosome (539,476 variants, geno_50KseqCHI19 scenario), the accuracy of the evaluations decreased by 4.8% on average compared with evaluations performed solely on 50k genotypes (geno_50K scenario). Two significant increases in accuracy were observed in the geno_50kseqCHI19 scenario for MY (+ 7.9%) and RUA (+3.6%). The highest decrease was observed for TO with a loss of 14.7% of accuracy, although not significant.

The additional selected variants on the updated version of the chip (geno_50kv2QTL) significantly increased the accuracy of genomic predictions for all the traits associated with the QTL region of chromosome 19 (FU, UFP, RUA, LSCS, MY, FY) except for PY. The mean gain was 3.4% for all traits considered. The highest gain was observed for FY (+10.0%) whereas the greatest loss was observed for PY (-7.4%). On the other hand, selecting randomly 178 variants on chromosome 19 did not have a significant impact on the accuracy of genomic evaluations for any of the traits evaluated except for FU which was marginally decreased by -0.1%.

In conclusion, geno_50kseqQTL outperformed all the other scenarios for the traits associated with the QTL region while not significantly decreasing the accuracy of evaluation for the other traits. The use of variants located in the QTL region systematically increased the accuracy of predictions compared with the same number of variants spread over the whole chromosome.

345 *WssGBLUP*

346 Table 3 shows the results for WssGBLUP (one weight for each SNP) for the different scenarios
347 tested.

348 When performing WssGBLUP on 50k genotypes (geno_50k), the gain in accuracy was 0.3%
349 on average for all traits, ranging from -14.6 % (LSCS; not significant) to +11.5% (MY).
350 Including the markers selected for the chip update (geno_50kv2QTL) led to an average gain in
351 accuracy of 2.9% with a significant decrease for LSCS (-14.1%) and a significant increase for
352 production traits, especially FY (+15.5%). The mean gain when using 50k genotypes and
353 sequence variants of the QTL region (geno_50kseqQTL) was 2.1% ranging from -15.8%
354 (LSCS) to +15.9% (FY).

355 In conclusion, WssGBLUP does not significantly improve accuracies compared with ssGBLUP
356 models. This method is only beneficial for production traits which were also improved when
357 using ssGBLUP.

358 *WssGBLUP$_{windows}$*

359 Every scenario using either 2.4-Mb or 40-SNP windows gave similar results (Table 4). When
360 using the updated version of the chip (geno_50kv2QTL), the mean gain in accuracy was 6.4%
361 and 6.3% for the 2.4-Mb and 40-SNP windows, respectively. In this scenario, the highest gain
362 was observed for MY with +19.0% and +19.2% for the 2.4-Mb and 40-SNP windows
363 respectively. The greatest losses were observed for LSCS with a significant decrease of -6.2%
364 and -7.1% for 2.4-Mb and 40-SNP windows, respectively.

365 Using all variants of the QTL region (geno_50kseqQTL) tended to be slightly less accurate with
366 an average gain of 4.2% and 4.5% for the 2.4-Mb and 40-SNP windows respectively. For this
367 scenario, the highest accuracies were observed for FY in both scenarios with increases

368 comprised between 14.9% and 18.7%. Highest losses were observed for LSCS with a decrease
369 comprised between -9.2% and -10.8%.

370 When the windows were located on the chromosome 19 only, the accuracy varied marginally
371 with an average decrease of -0.5%. The decrease was particularly significant for LSCS (-
372 12.4%), UP (-10.9%) and TO (-10.7%). MY was better predicted with an increase in accuracy
373 of 15.8%.

374 In conclusion, WssGBLUP$_{windows}$ does not improve genomic evaluations as well as a ssGBLUP
375 model except for production traits for which it results in high gains in accuracy.

376 # DISCUSSION

377 ## *Variants selection*

378 As a result of the efforts of the VarGoats Consortium, a large amount of sequence data is now
379 available to the research community. However, sequence data comprise vast numbers of
380 variants (23,337,436 for the whole genome after the filtering process). Careful variant selection
381 is therefore crucial to avoid burdening genomic evaluations with too much information. In this
382 study, we assessed various options for selecting variants. We are aware that the imputation
383 quality is lower than in other species which might lead to a degradation of predictions as
384 suggested by Perez-Enciso et al (Pérez-Enciso et al., 2015). However, the genomic predictions
385 were improved in our study. Further investigations will be needed once the update of the current
386 chip will be validated in order to confirm the improvement we observed with more reliable
387 genotypes. Indeed, it has previously been shown that when all variants on chromosome 19
388 (539,476 variants) are included in ssGBLUP, the genomic evaluations tend to be less accurate.
389 Consistent with this finding, we found that adding sequence data for the whole chromosome 19
390 introduced noise in the evaluations compared with genomic evaluations using chip data only.
391 Figure 3 shows the variant effects for the FU trait in the different scenarios. Among the traits

associated with the QTL region, FU is the most affected in the geno_50kseqCHI19 scenario (-12.6% accuracy). The introduction of sequence data of the whole CHI19 (geno_50kseqCHI19) led to a disruption of estimated effects on the whole genome (Figure 3). The distribution of the variant effects is completely modified including effects of the 50k SNP variants. The latter are greater than in any other of the tested scenarios (Figure 3). Besides, some of the signals observed do not match any of the previously identified QTL regions and might have introduced errors in the genomic prediction equations. Similar patterns were observed for every trait showing decreased evaluation accuracy when sequence variants covering the whole chromosome were included.

In ssGBLUP, when the selection of variants is limited to the QTL region (between 24.72 and 28.38 Mb, 22,269 sequence variants selected), the gains in accuracy compared with the geno_50k scenario were high for traits associated with the region without a loss in accuracy for the other traits (TO and UP). The mean percentage of explained variance for the region between 23 and 30 Mb was 1.77% with the 50k genotypes alone but increased to 9.37% on average when sequence variants in the QTL region were added. In ssGBLUP, this increased percentage led to tremendous gains in accuracy especially for MY (+13,7%), PY (+12.5%), FY (17.9%) and UFP (+9.0%) for which the variance explained by the region between 23 and 30 Mb reached respectively 8.24%, 9.03%, 7.30% and 17.77% (Figure 2). However, we also observed in the geno_50kseqQTL scenario an increase in the percentage of variance explained by this region for traits that are not associated with this region (TO and UP): +4.1% and +4.6% for TO and UP, respectively. This increase might be an artefact linked to the enrichment of the area, each SNP having a small effect. This artefact could explain the deterioration of accuracy for these traits in this scenario. Besides, with the inclusion of sequence data, we introduce variants which are not completely independent leading to an over- or under-estimation of the variance they explain. In light of these findings, the updated chip appears as the best scenario as it has a lesser

417 impact on the traits not associated with CHI19 than the sequence data of the QTL region (Figure
418 2) and as the choice of the markers partly took into account their LD.

419 Selecting variants within QTL regions therefore represents a great opportunity to improve
420 routine genomic evaluations without losing in speed. Identifying the causal mutation for each
421 trait might lead to further increases in the accuracy of genomic evaluations. However
422 pinpointing one variant is difficult given the number of variants in the area, the high linkage
423 disequilibrium and the low estimated recombination rates (lower than 1) (Rachel Rupp, INRAE,
424 personal communication). Nevertheless, Bolormaa *et al* (Bolormaa et al., 2019) reported a
425 significant impact of the accuracy of imputation to sequence level on the accuracy of genomic
426 prediction in sheep data. It therefore seems likely that sequencing more individuals could
427 further improve evaluations by lifting imputation errors and refining the QTL region.

428 Nevertheless, it is important to bear in mind that imputation has to be performed to get sequence
429 information for every genotyped individual. This is a time-consuming process and imputed
430 sequences represent a large amount of data that require significant storage capacities. In our
431 study, we also show that the future version of the Illumina GoatSNP50 BeadChip is promising
432 if the additional variants selected in the QTL region (178 SNP variants) are added for routine
433 genomic evaluations. Indeed, the mean gain of accuracy was 4.9% with ssGBLUP and ranged
434 from -0.1% (TO and LSCS; not significant) to 10.1% (FY). This updated chip has an increased
435 density within the QTL region of chromosome 19 with 308 SNPs (130 SNPs from version 1 +
436 178 new selected variants) located between 23 and 30 Mb with an expected average spacing of
437 22,675 bp (ranging from 4 to 166,413 bp). However, the selected variants to be added to the
438 chip have yet to be validated with a cluster file. A confirmation study will be needed with the
439 real genotypes and the variants still present in the genotypes after the quality check before
440 implementation in routine evaluation pipelines.

Similar studies were performed by VanRaden *et al* (VanRaden et al., 2017) on Holstein bulls. They achieved less significant gains in accuracy when sequence data was added to HD genotypes with only 0.6 percentage points when using solely SNPs and 0.4 percentage points when adding both SNPs and insertion/deletions (indels). This might be due to the fact that HD genotypes are already exhaustive compared with 50k genotypes so gains might be lower. In goats, the only genotyping tool available is a medium density chip, so it makes sense that the gains in accuracy when sequence data is added are higher than in dairy cattle. In the aforementioned study, adding 16,648 candidate SNPs to the routinely used 60k SNPs systematically resulted in an average gain that was higher for all traits. However, the average increase was marginal and other studies also showed little gain when adding sequence information to 50k genotypes in cattle (B. O. Fragomeni et al., 2019).

Several studies have demonstrated that selecting a subset of variants to be included in the genomic evaluations is a particularly relevant approach. VanRaden *et al* (VanRaden et al., 2017) demonstrated that selecting a subset of SNPs (nearly 17,000 SNPs) with the highest effects maximized the gains in accuracy (2.7 percentage points). Brøndum *et al* (Brøndum et al., 2015) selected 1,623 variants for the custom Illumina BovineLD SNP chip for Nordic breeds by performing association. They observed gains in accuracy ranging from 3 to 5 percentage points for production traits, less than 1 percentage point for udder health and 0.5 percentage point for fertility. However, these gains are small compared with our findings. On simulated data, including QTL information seemed to increase the accuracy of the evaluations similarly to our results in the Saanen breed. Fragomeni *et al* (Fragomeni et al., 2017) simulated livestock populations and obtained a gain of 8.2% when unweighted QTL information was added to a ssGBLUP model.

464 *Models comparisons*

465 Our study follows the work of Teissier *et al* (Teissier et al., 2019; Teissier et al., 2018) who
466 improved the accuracy of genomic evaluations of protein content by 4% when using
467 WssGBLUP models taking the highest effect for a 40-SNP window. They also reported
468 significant gains for milk yield (+7 percentage points), fat and protein yields (+4 and +5
469 percentage points, respectively), udder floor position (+4 percentage points), rear udder
470 attachment (+1 percentage point) in Saanen goats compared with a ssGBLUP model. Our
471 results show that using a standard WssGBLUP including sequence data from the QTL region
472 slightly increased the accuracy of genomic evaluations (+2.1% on average) compared with
473 ssGBLUP performed on 50k genotypes. This is especially true for the production traits MY,
474 FY, and PY for which the gains in accuracy were always higher than for other traits. This has
475 also been reported in previous studies on 50k genotypes (Teissier et al., 2019). However, as
476 previously observed by Teissier *et al* (Teissier et al., 2019; Teissier et al., 2018), the increase
477 in gain is variable depending on the trait. Indeed, the accuracies for production traits are
478 increased by 15% on average whereas other traits tended to show slight decreases. LSCS was
479 the most impacted trait with an evaluation accuracy that decreased by -15.8%. This might be
480 related to the variant selection process for the update which selected variants significant for the
481 highest number of traits but without checking the number of variants associated with each trait.
482 Table 5 shows the representativeness of each trait with the variants selected for the update.
483 LSCS is currently not represented on the updated version of the chip because of its low
484 association to the QTL region. This might explain why the selected variants tend to introduce
485 noise in the evaluations for this particular trait, and adding weights to the variants within the
486 region might have amplified the phenomenon. Nevertheless, we reasonably question whether
487 the ratio between the gain in accuracy and the time spent preparing the data (quality control,

488 imputation, data formatting for software, etc.) is significant enough to implement the procedure
489 on a routine basis.

490 In the WssGBLUP$_{windows}$ models, regardless of the method used to build the windows, the QTL
491 information tended to be smoothed (Figure 4). All weighted scenarios using sequence
492 information from the QTL region (geno_50kseqQTL) led to significant decreases in the
493 accuracy of genomic evaluations for both type traits (except for UFP and RUA) and LSCS
494 compared with simple ssGBLUP on 50k genotypes. This finding might be caused by the use of
495 extremely summarized information. Indeed, when windows were built on a distance basis (2.4
496 Mb), the whole chromosome 19 was spanned by only 26 windows and windows of 40
497 consecutive variants led to the construction of 586 windows on chromosome 19. These results
498 are inconsistent with previous results in the French Saanen breed. Indeed, Teissier *et al* (Teissier
499 et al., 2018) improved genomic evaluations of the protein content when using windows of 40
500 consecutive variants. However, the casein region is smaller and has larger effects on the trait
501 they studied in Saanen goats than chromosome 19, as shown by Carillier-Jacquin et al.
502 (Carillier-Jacquin et al., 2016) who reported that the genetic variance explained by the αs1-
503 casein gene reached 38% in this breed. The 50k SNPs in the QTL region on chromosome 19
504 only explain between 6.6% and 21.5% of the genetic variance, depending on the trait. Its total
505 effect might therefore be diluted when the region is divided into windows.

506 Regardless of the method used to build the windows, the variants selected for the chip update
507 in the QTL region led to an average gain in accuracy (6.4% for 2.4-Mb windows and 6.3% for
508 40-variant windows). The updated version of the chip therefore seems to perform slightly better
509 than 50k genotypes (+ 2.2% on average) or the use of sequence information for the QTL region
510 (+4.2% on average). This result is promising; however, the gain is variable depending on the
511 trait. Production traits are 17.7% more accurate when using WssGBLUP$_{windows}$. Other traits
512 tended to show decreases in accuracy except for UFP (+8%). Similarly to WssGBLUP, the

513 traits that showed the highest gains with the 50k chip update in WssGBLUP$_{windows}$ models were
514 those with the highest number of associated variants in the update (Table 5).

515 The aim of this study was to investigate whether sequence data could be included in routine
516 genomic evaluations and if so which model would provide optimal results. The findings we
517 present here are preliminary and further investigation is needed to assess the influence of other
518 known QTL regions in French dairy goats, for example the DGAT region (CHI14) for fat
519 content and the casein region (CHI6) for protein content. Different models have already been
520 studied in French dairy goats using solely 50k genotypes (Teissier et al., 2019; Teissier et al.,
521 2018) and in the light of our results the availability of sequence data represents a great
522 perspective. When choosing the optimal strategy, one should bear in mind that production traits
523 have a higher weight in the calculation of indexes than type traits (Larroque et al., 2011).
524 Indeed, the formula for the Saanen breed is as follows:

525 $$Combined\ Index = Production\ Index + 0.6 \times Morphology\ Index$$

526 An evaluation model that tends to improve the accuracy of prediction of production traits might
527 therefore be preferable. It hence seems reasonable to exclude using ssGBLUP with the updated
528 version of the chip as it significantly deteriorates the accuracy of evaluation of PY (-7.4%).
529 However, a ssGBLUP model using sequence data from the QTL seems appropriate. All
530 WssGBLUP and WssGBLUP$_{windows}$ scenarios also increased the accuracy of prediction of
531 production traits.

532 Even though type traits are of smaller importance in the combined index, scenarios and models
533 that do not deteriorate the prediction of type traits should be favored. In the Morphology Index,
534 each type trait is assigned the same weight. In this light, a WssGBLUP model with windows of
535 2.4 Mb on the updated version of the chip seems the best option as it is the model that causes

536 the smallest decrease in the accuracy of prediction for LSCS, UP and TO while improving the
537 accuracy of genomic predictions for other traits.

538 CONCLUSIONS

539 Including sequence data from the QTL region of chromosome 19 led to significant gains in
540 accuracy for the genomic evaluation of traits associated with this region of the genome. The
541 most time-efficient way to take sequence data into account on a routine basis seems to be a
542 simple ssGBLUP model using variants of the QTL region. However, the upcoming chip update
543 paves the way for a more strategic approach. Indeed, information from only a few carefully
544 selected variants led to increased accuracies for the traits of interest. If production traits are to
545 be emphasized, the WssGBLUP model makes the most of the updated chip with significant
546 increases for MY, PY and FY. Besides, this update is interesting as it would avoid time-
547 consuming imputation and data formatting processes while providing reliable genotypes.

548 REFERENCES

549 Bolormaa, S., Chamberlain, A. J., Khansefid, M., Stothard, P., Swan, A. A., Mason, B., Prowse-
550 Wilkins, C. P., Duijvesteijn, N., Moghaddar, N., van der Werf, J. H., Daetwyler, H. D., &
551 MacLeod, I. M. (2019). Accuracy of imputation to whole-genome sequence in sheep.
552 *Genetics Selection Evolution*, *51*(1), 1. https://doi.org/10.1186/s12711-018-0443-5

553 Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S.
554 (2015). Quantitative trait loci markers derived from whole genome sequence data
555 increases the reliability of genomic prediction. *Journal of Dairy Science*, *98*(6), 4107–
556 4116. https://doi.org/10.3168/jds.2014-9005

557 Carillier-Jacquin, C., Larroque, H., & Robert-Granié, C. (2016). Including α s1 casein gene
558 information in genomic evaluations of French dairy goats. *Genetics Selection Evolution*,

559 *48*(1), 1–13. https://doi.org/10.1186/s12711-016-0233-x

560 Carillier, C., Larroque, H., Palhière, I., Clément, V., Rupp, R., & Robert-Granié, C. (2013). A first
561 step toward genomic selection in the multi-breed French dairy goat population. *Journal*
562 *of Dairy Science*, *96*(11), 7294–7305. https://doi.org/10.3168/jds.2013-6789

563 Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not
564 genotyped. *Genetics Selection Evolution*, *42*(3), 1–8. https://doi.org/10.1186/1297-9686-
565 42-2

566 Clément, V., Martin, P., & Barillet, F. (2006). Elaboration d ' un index synthétique caprin
567 combinant les caractères laitiers et des caractères de morphologie mammaire
568 Elaboration of a total merit index combining dairy and udder type traits. *Renc. Rech.*
569 *Ruminants*, *1*, 209–212.

570 Fragomeni, B. O., Lourenco, D. A. L., Legarra, A., VanRaden, P. M., & Misztal, I. (2019).
571 Alternative SNP weighting for single-step genomic best linear unbiased predictor
572 evaluation of stature in US Holsteins in the presence of selected sequence variants.
573 *Journal of Dairy Science*, *102*(11), 10012–10019. https://doi.org/10.3168/jds.2019-16262

574 Fragomeni, Breno O., Lourenco, D. A. L., Masuda, Y., Legarra, A., & Misztal, I. (2017).
575 Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genetics*
576 *Selection Evolution*, *49*(1), 1–11. https://doi.org/10.1186/s12711-017-0335-0

577 Hayes, B. J., Macleod, I. M., Daetwyler, H. D., Bowman, P. J., Chamberlian, A. J., Vander Jagt,
578 C. J., Capitan, A., Pausch, H., Stothard, P., Liao, X., Schrooten, C., Mullaart, E., Fries, R.,
579 Guldbrandtsen, B., Lund, M. S., Boichard, D. A., Veerkamp, R. F., Vantassell, C. P., Gredler,
580 B., … Goddard, M. E. (2014). Genomic Prediction from Whole Genome Sequence in

581 Livestock: the 1000 Bull Genomes Project. *Proceedings, 10th World Congress of Genetics*
582 *Applied to Livestock Production*.

583 Hickey, J. M., Kinghorn, B. P., Tier, B., Van Der Werf, J. H. J., & Cleveland, M. A. (2012). A
584 phasing and imputation method for pedigreed populations that results in a single-stage
585 genomic evaluation. *Genetics Selection Evolution*, *44*(1), 1–11.
586 https://doi.org/10.1186/1297-9686-44-9

587 Larroque, H., Astruc, J. M., Barbat, A., Barillet, F., Boichard, D., Bonaïti, B., Clément, V., David,
588 I., Lagriffoul, G., Palhière, I., Piacère, A., Robert-Granié, C., & Rupp, R. (2011). National
589 genetic evaluations in dairy sheep and goats in France. *Proceedings of the 62nd Annual*
590 *Meeting of the European Federation of Animal Science*.

591 Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and
592 genomic information. *Journal of Dairy Science*, *92*(9), 4656–4663.
593 https://doi.org/10.3168/jds.2009-2061

594 Martin, P, Palhière, I., Maroteau, C., Clément, V., David, I., Tosser-Klopp, G., & Rupp, R. (2018).
595 Genome-wide association mapping for type and mammary health traits in French dairy
596 goats identifies a pleiotropic region on chromosome 19 in the Saanen breed. *Journal of*
597 *Dairy Science*, *0*(0), 5214–5226. https://doi.org/10.3168/jds.2017-13625

598 Martin, Pauline, Palhière, I., Maroteau, C., Bardou, P., Canale-Tabet, K., Sarry, J., Woloszyn, F.,
599 Bertrand-Michel, J., Racke, I., Besir, H., Rupp, R., & Tosser-Klopp, G. (2017). A genome
600 scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing
601 milk fat content. *Scientific Reports*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-
602 02052-0

603 Misztal, I., Tsuruta, S., Strabel, T., Auvrey, B., Druet, T., & Lee, D. (2002a). BLUPF90 and related
604 programs. *Proceedings of the 7th World Congress on Genetics Applied to Livestock*
605 *Production: 19–23 August 20*.

606 Misztal, I., Tsuruta, S., Strabel, T., Auvrey, B., Druet, T., & Lee, D. (2002b). BLUPF90 and related
607 programs. *Proeedings of the 7th World Congress on Genetics Applied to Livestock*
608 *Production: 19-23 August 2002*.

609 Moghaddar, N., Macleod, I. M., Duijvesteijn, N., Bolormaa, S., Khansefid, M., Swan, A. A.,
610 Daetwyler, H. D., & van der Werf, J. H. J. (2018). Genomic evaluation based on selected
611 variants from imputed whole-genome sequence data in Australian sheep populations.
612 *Proceedings of the World Congress on Genetics Applied to Livestock Production*.

613 Mucha, S., Mrode, R., Coffey, M., Kizilaslan, M., Desire, S., & Conington, J. (2017). Genome-
614 wide association study of conformation and milk yield in mixed-breed dairy goats. *Journal*
615 *of Dairy Science*, *101*(3), 2213–2225. https://doi.org/10.3168/jds.2017-12919

616 Pérez-Enciso, M., Rincón, J. C., & Legarra, A. (2015). Sequence- vs. chip-assisted genomic
617 selection: Accurate biological information is advised. *Genetics Selection Evolution*, *47*(1),
618 1–14. https://doi.org/10.1186/s12711-015-0117-5

619 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar,
620 P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome
621 association and population-based linkage analyses. *American Journal of Human Genetics*,
622 *81*(3), 559–575. https://doi.org/10.1086/519795

623 Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype
624 imputation using information from relatives. *BMC Genomics*, *15*(1).

https://doi.org/10.1186/1471-2164-15-478

Talouarn, E., Bardou, P., Palhière, I., Oget, C., Clément, V., The VarGoats Consortium, Tosser-Klopp, G., Rupp, R., & Robert-Granié, C. (2020). Genome wide association analysis on semen volume and milk yield using different strategies of imputation to whole genome sequence in French dairy goats. *BMC Genetics*, *21*(1), 1–13. https://doi.org/10.1186/s12863-020-0826-9

Teissier, M., Larroque, H., & Robert-Granie, C. (2019). Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats. *Journal of Dairy Science*, *102*(4), 3142–3154. https://doi.org/10.3168/jds.2018-15650

Teissier, Marc, Larroque, H., & Robert-Granié, C. (2018). Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: A quantitative trait influenced by a major gene. *Genetics Selection Evolution*, *50*(1), 1–12. https://doi.org/10.1186/s12711-018-0400-3

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., Dong, Y., Donnadieu-Tonon, C., Eggen, A., Heuven, H. C. M., Jamli, S., Jiken, A. J., Klopp, C., Lawley, C. T., McEwan, J., Martin, P., Moreno, C. R., Mulsant, P., Nabihoudine, I., Pailhoux, E., … Zhao, S. (2014). Design and characterization of a 52K SNP chip for goats. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0086227

VanRaden, P. M., & Wiggans, G. R. (1991). Derivation, Calculation, and Use of National Animal Model Information. *Journal of Dairy Science*, *74*(8), 2737–2746. https://doi.org/10.3168/jds.S0022-0302(91)78453-1

VanRaden, Paul M., Tooker, M. E., O'Connell, J. R., Cole, J. B., & Bickhart, D. M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genetics Selection Evolution*, *49*(1), 1–12. https://doi.org/10.1186/s12711-017-0307-4

Wang, H., Misztal, I., Aguilar, I., Legarra, A., & Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genetics Research*, *94*(2), 73–83. https://doi.org/10.1017/S0016672312000274

Wang, Huiyu, Misztal, I., Aguilar, I., Legarra, A., Fernando, R. L., Vitezica, Z., Okimoto, R., Wing, T., Hawken, R., & Muir, W. M. (2014). Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Frontiers in Genetics*, *5*(MAY), 1–10. https://doi.org/10.3389/fgene.2014.00134

Williams, E. J. (1959). The Comparison of Regression Variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 396–399. https://doi.org/10.1111/j.2517-6161.1959.tb00346.x

Zhang, X., Lourenco, D., Aguilar, I., Legarra, A., & Misztal, I. (2016). Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS. *Frontiers in Genetics*, *7*(AUG), 1–14. https://doi.org/10.3389/fgene.2016.00151

Acknowledgements

This study would not have been possible without the sequence data provided by the VarGoats Consortium (http://www.goatgenome.org/ vargoats.html) and previous work by the International Goat Genome Consortium (IGGC, http://www.goatgenome.org/) and ADAPTmap Consortium (http://www.goatadaptmap.org/) providing relevant DNA samples,

669 genotyping tools and genotyping data through their collaborative networks. Genotypes were
670 funded by several projects: the French Genovicap and Phenofinlait programmes (ANR, Apis-
671 Gène, CASDAR, FranceAgriMer, France Génétique Elevage, the French Ministry of
672 Agriculture Agrifood, and Forestry), the European 3SR project, and Maxi'male (CASDAR).
673 We also would like to thank the CapGenes breeding organization for the data provided. We are
674 grateful to the Genotoul bioinformatics platform Toulouse MidiPyrénées and the CTIG (Centre
675 de Traitement de l'Information Génétique) of INRAE Jouy-en-Josas for providing computing
676 resources. The authors thank Ignacy Misztal (University of Georgia, USA) for the blup90iod2
677 program. The first author also received financial support from the Occitanie region and the
678 French Research National Research Institute for Agriculture, Food and Environment (INRAE
679 – Animal Genetic division).

680 Authors'contribution

681 CRG and RR designed the study. ET analyzed the data and drafted the manuscript. PB called
682 the variants and provided support in computing. MT helped with the implementation of the
683 different genomic models. HL, IP and VC provided information on the current routine
684 evaluations. IP provided part of the performance file and chose individuals to be sequenced.
685 ET, GTK, CRG and RR interpreted the results. RR and CRG improved the manuscript. All
686 authors read and approved the final manuscript.

687 Competing interests

688 The authors declare they do not have any competing interests.

689 Abbreviations

690 DYD: Daughter Yield Deviation

691 FU: Fore Udder

692 FY: Fat Yield

693 GBLUP: Genomic Best Linear Unbiased Prediction

694 GEBV: Genomic Estimated Breeding value

695 HD: High-Density

696 LSCS: Somatic Cells Score

697 MAF: Minor Allele Frequency

698 MY: Milk Yield

699 PY: Protein Yield

700 QTL: Quantitative Trait Loci

701 REML: REstricted Maximum Likelihood

702 RUA: Rear Udder Attachment

703 SD: Standard Deviation

704 ssGBLUP: single-step GBLUP

705 TO: Teat Orientation

706 UFP: Udder Floor Position

707 UP: Udder Profile

708 WssGBLUP: Weighted single-step GBLUP

Tables

Table 1. Summary statistics for the traits used in the genetic evaluations of French Saanen goats (born between 1980 and 2017)

Trait	*Number of lactations*[1]	*Number of females with phenotypes*	*Min*	*Mean*	*SD*	*Max*
MY (kg)	3,470,255	1,242,020	34.51	837.05	266.03	2,615.04
PY (kg)	3,470,255	1,274,581	1.27	25.05	8.08	84.64
FY (kg)	3,470,255	1,271,383	1.09	28.06	10.01	111.92
LSCS	1,449,698	705,753	-0.58	8.82	1.34	13.57
FU	-	160,086	1	3.29	1.16	9
TO	-	160,086	1	4.03	0.86	9
UFP	-	160,086	1	6.23	1.14	9
UP	-	160,086	1	6.28	1.34	9
RUA	-	160,086	1	4.83	1.67	9

SD: Standard Deviation, MY: Milk Yield, PY: Protein Yield, FY: Fat Yield, LSCS: Somatic Cell Score, FU: Fore Udder, TO: Teat Orientation, UFP: Udder Floor Position, UP, Udder Profile, RUA: Rear Udder Attachment.

[1] Measurements for type traits were performed once in each animal's lifetime

717 **Table 2.** Summary of the scenarios tested using ssGBLUP

Name of the scenario	*Variants included in the scenario*					*Total number of variants included in the scenario*
	50k markers	178 SNPs in chromosome 19 QTL region on the chip update	178 randomly selected SNPs on chromosome 19	539,476 sequence variants over the whole chromosome 19[1]	22,269 sequence variants in the QTL region of chromosome 19[1]	
Ref *geno_50k*	x					47,147
1 *geno_50kv2QTL*	x	x				47,325
2 *geno_50kv2_random*	x		x			47,325
3 *geno_50kseqCHI19*	x			x		586,623
4 *geno_50kseqQTL*	x				x	69,416

718 [1] variants obtained on imputed sequence of 1,207 Saanen individuals

719 **Table 3.** Accuracy of genomic predictions using standard WssGBLUP in the validation
720 population of 146 Saanen bucks

Trait	*geno_50k*	*geno_50v2QTL*	*geno_50kseqQTL*

FU	0.60*	0.62*	0.60*
TO	0.48*	0.48	0.48
UFP	0.66*	0.66*	0.67*
UP	0.35*	0.35*	0.35*
RUA	0.63	0.64	0.63
LSCS	0.42	0.43*	0.42*
MY	0.54*	0.57*	0.56*
FY	0.47*	0.50*	0.50*
PY	0.50*	0.53*	0.53*

721 * Significantly different from the correlation obtained with ssGBLUP using solely 50k
722 genotypes

723 FU: fore udder, FY: Fat Yield, LSCS: Somatic cell score, MY: Milk Yield, PY: Protein Yield,
724 RUA: Rear Udder Attachment. TO: teat orientation, UFP: Udder Floor Position, UP, Udder
725 profile

726 **Table 4.** Accuracy of genomic predictions using WssGBLUP with different windows in the
727 validation population of 146 Saanen goats

scenario	*geno_50k*	*geno_50kv2QTL*		*geno_50kseqQTL*		
windows	2.4 Mb (≈ 40 SNP)	40 SNP	2.4Mb	40 SNP	2.4 Mb	2.4Mb on CHI19 only
FU	0.62*	0.64	0.64	0.63	0.63	0.57*
TO	0.48*	0.48*	0.48*	0.47*	0.47*	0.44*
UFP	0.67*	0.69*	0.68*	0.68*	0.67*	0.66*
UP	0.38	0.38	0.38	0.37	0.37	0.34*
RUA	0.65*	0.67*	0.67*	0.66*	0.65	0.65
LSCS	0.48	0.46*	0.46*	0.45*	0.44*	0.43*
MY	0.53*	0.58*	0.58*	0.56*	0.56*	0.54*
FY	0.44*	0.50*	0.50*	0.50*	0.51*	0.50*
PY	0.49*	0.54*	0.54*	0.52*	0.52*	0.49

728 * Significantly different from the correlation obtained with ssGBLUP using solely 50k
729 genotypes

730 FU: fore udder, FY: Fat Yield, LSCS: Somatic cell score, MY: Milk Yield, PY: Protein Yield,
731 RUA: Rear Udder Attachment. TO: teat orientation, UFP: Udder Floor Position, UP, Udder
732 profile

733 **Table 5.** Representativeness of each trait associated with the QTL region of CHI19 among the
734 178 variants selected for the chip update

Trait associated to the QTL region of chromosome 19	*Number of variants selected for the update associated with the trait*
FU	12
UFP	150
RUA	52
LSCS	0
MY	104
FY	37
PY	62

735 MY: Milk Yield. PY: Protein Yield, FY: Fat Yield, LSCS: Somatic cell score, FU: fore udder,
736 UFP: Udder Floor Position, RUA: Rear Udder Attachment

737 ## Figures

738

739

740

741

742

743

744

745

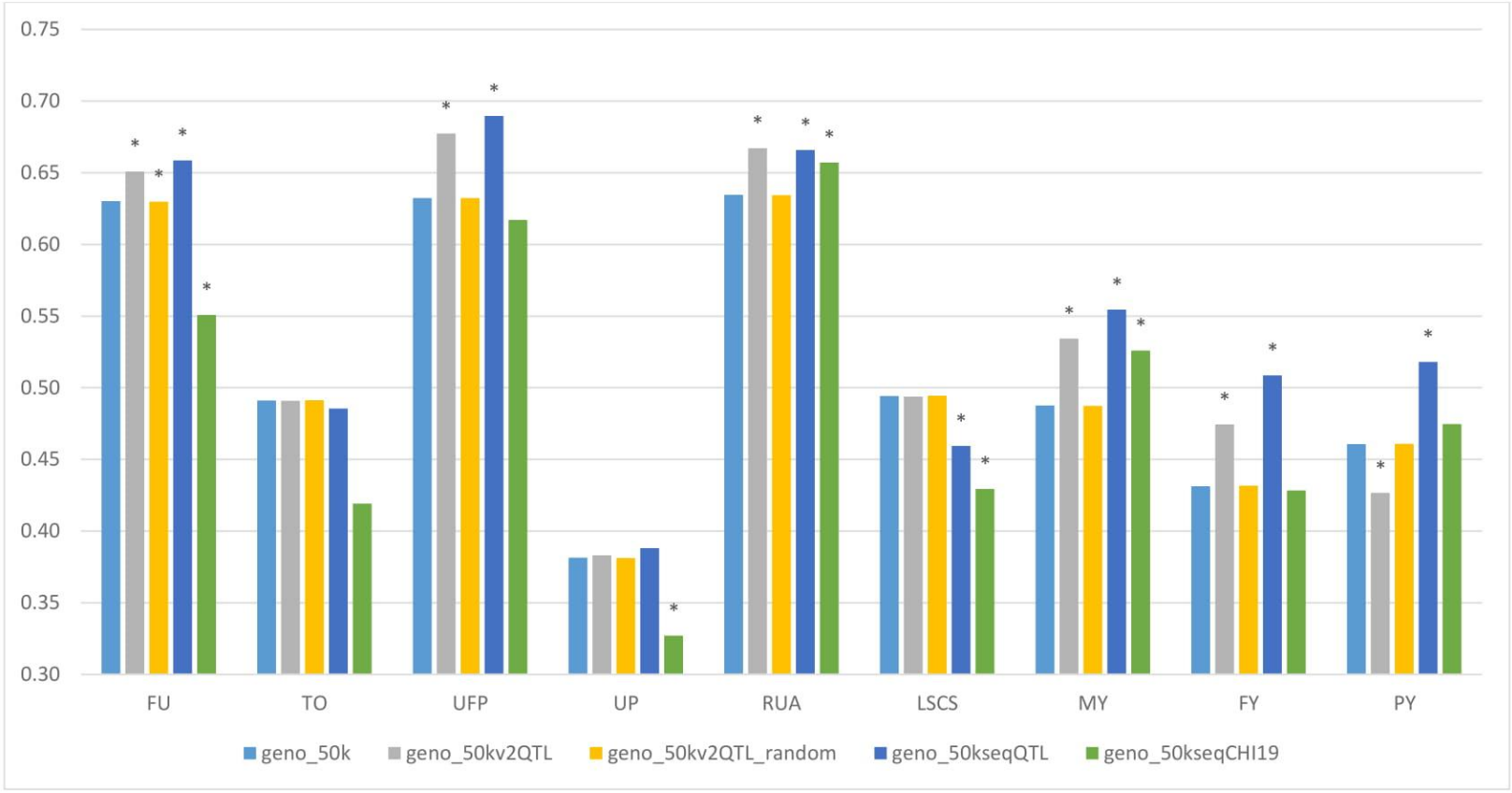
746

747

748

749

750 **Figure 1.** Accuracies of genomic predictions for ssGBLUP model on different scenarios in the 146 validation Saanen individuals



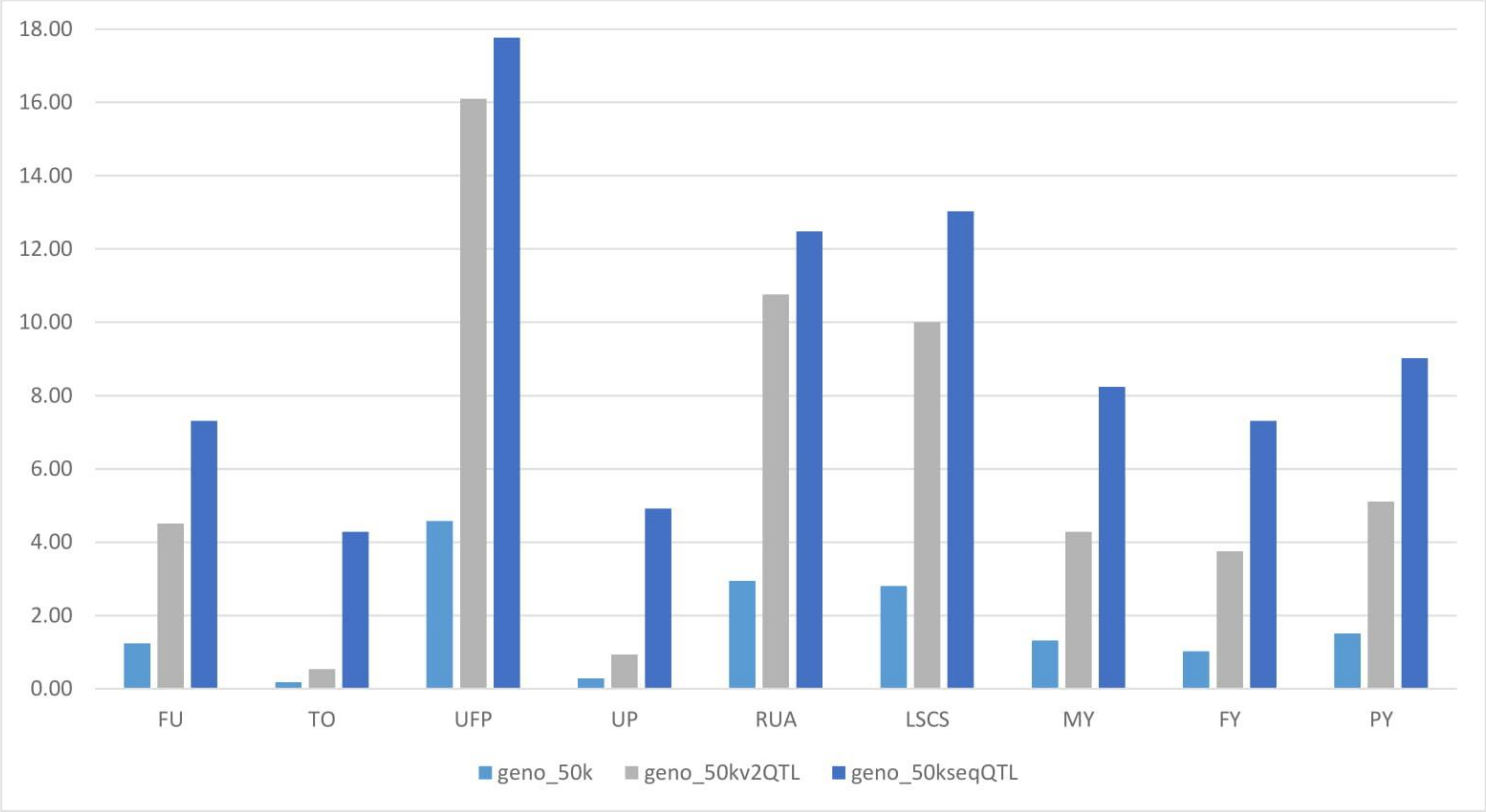
751

752 *: significantly different from accuracy obtained in a ssGBLUP using 50k genotypes

753 FU: fore udder, FY: Fat Yield, LSCS: Somatic cell score, MY: Milk Yield, PY: Protein Yield, RUA: Rear Udder Attachment. TO: teat orientation,

754 UFP: Udder Floor Position, UP, Udder profile

Figure 2. Percentage of the variance explained by the region between 23 and 30Mb of CHI19 for each trait in the geno_50k (130 variants), geno_50kv2QTL (308 variants) and geno_50kseqQTL (22,399 variants) scenarios



FU: fore udder, FY: Fat Yield, LSCS: Somatic cell score, MY: Milk Yield, PY: Protein Yield, RUA: Rear Udder Attachment. TO: teat orientation, UFP: Udder Floor Position, UP, Udder profile

Figure 3. Variant effects in the ssGBLUP scenarios tested for fore udder trait in the 146 validation Saanen individuals 50k markers on the left of the red line; additional variants on the right side of the red line

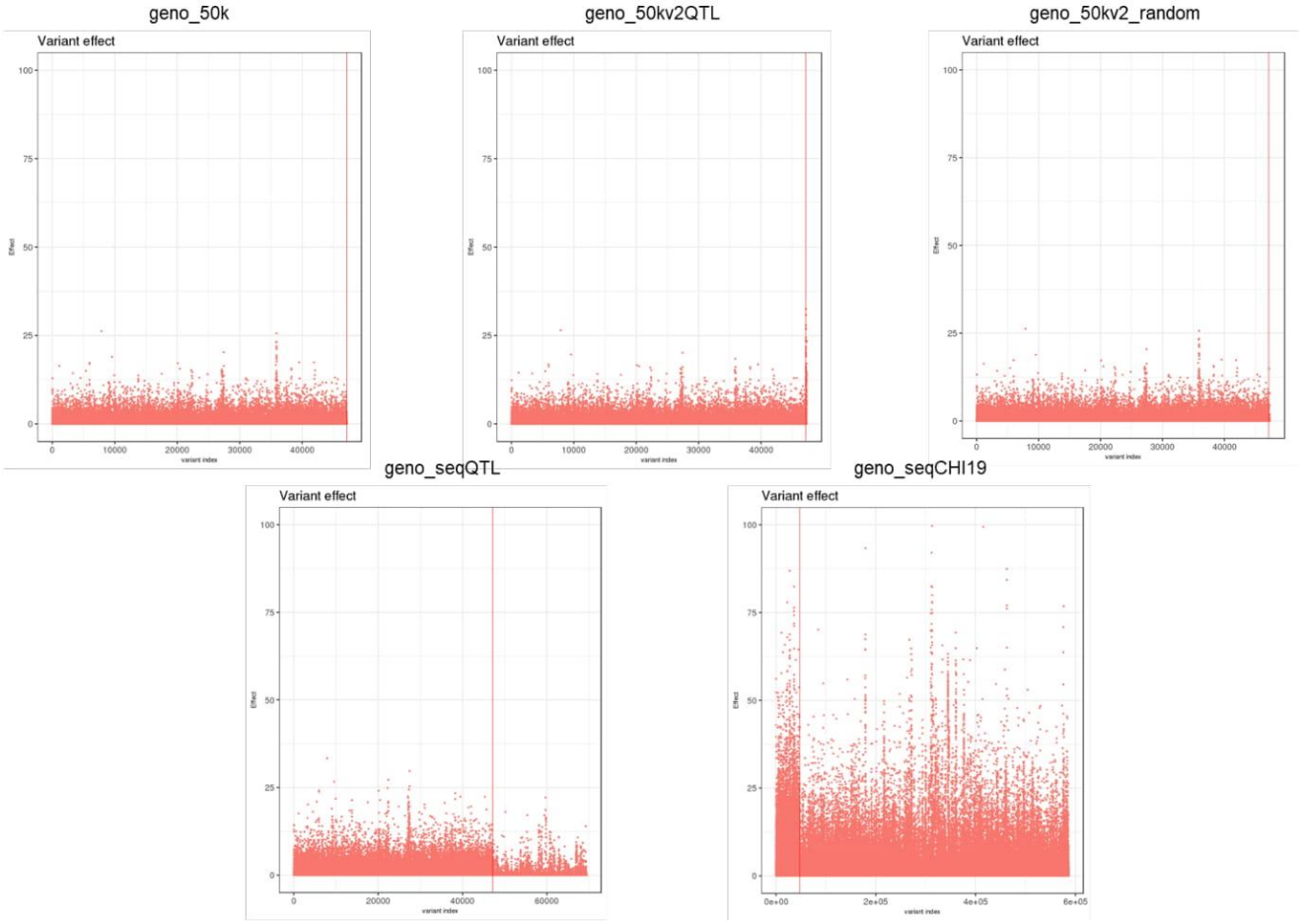
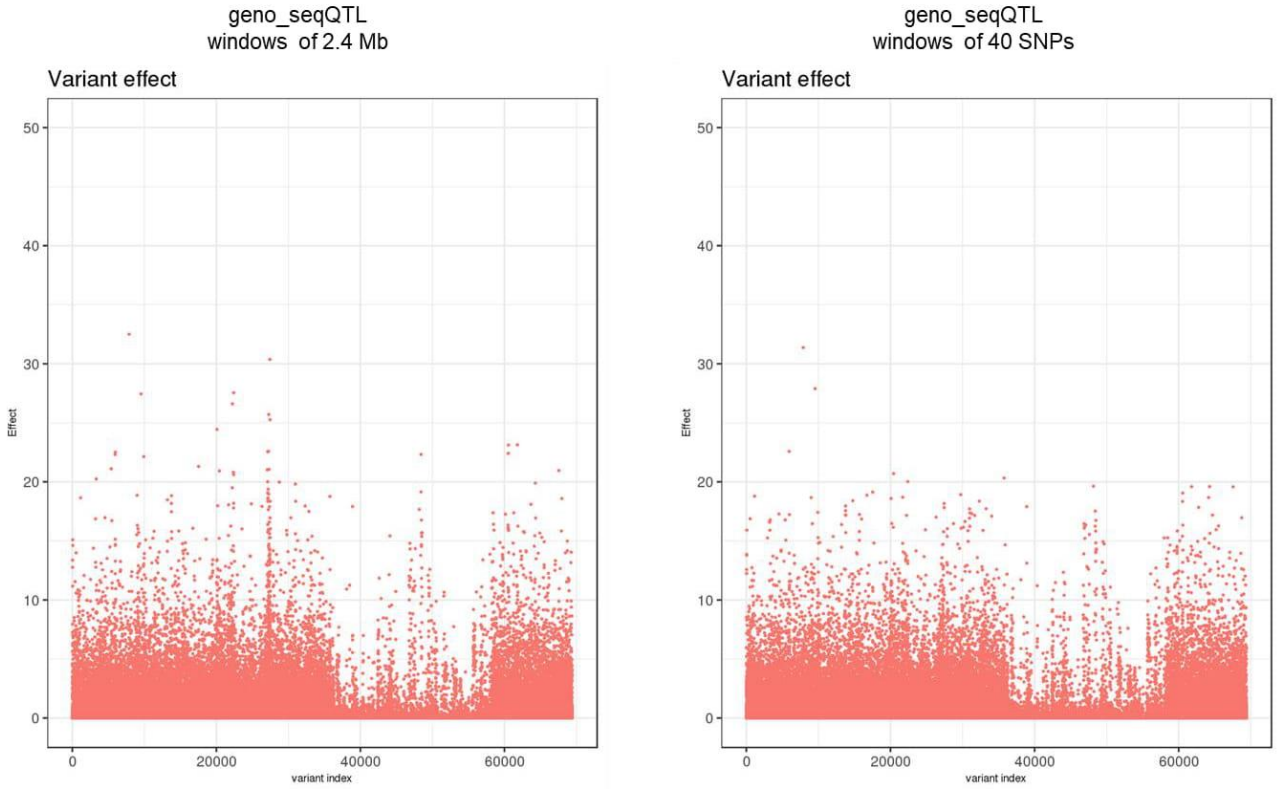


Figure 4. Variant effects in the WssGBLUP scenarios using sequence variants of the QTL region for fore udder trait in the 146 validation Saanen individuals



I.3. Utilisation de l'information de « groupe » dans les évaluations génomiques

A l'aide d'ACP effectuées sur les variants de séquence du chromosome 19 entre 23 et 30Mb, nous avons identifié au chapitre 3 trois profils génomiques de Saanen françaises. Ils peuvent être distingués à l'aide d'un ensemble de variants qui recouvrent une région de 3,6 Mb dans la région QTL du chromosome 19. Trois marqueurs de la puce 50kv1 ont été retenus car ils distinguaient correctement les trois groupes tels que définis précédemment. Pour qu'un maximum d'individus Saanen avec des génotypes 50k soit ainsi classés dans un groupe, nous avons considéré qu'un individu appartenait au groupe s'il possédait les génotypes attendus pour au moins 2 des 3 marqueurs.

Pour les évaluations génomiques, l'information de groupe a été encodée sous la forme d'un pseudo-SNP puis ajouté aux génotypes 50k utilisés pour l'évaluation. Ainsi le « génotype » de ce pseudo-SNP prend la valeur 2 lorsque l'individu est du groupe 1, 1 pour un individu du groupe 2 et 0 pour un individu du groupe 3. Les 15 individus à qui il a été impossible d'attribuer un groupe ont un génotype inconnu (codé avec la valeur 5) pour ce pseudo-SNP.

Plusieurs scenarios ont été envisagés : (1) l'addition simple de l'information de groupe en tant que pseudo-SNP à un génotype 50k (geno_groupe) (2) l'ajout du pseudo-SNP et le retrait dans les génotypes 50k des SNP de la région QTL (geno_groupeseul) (3) le retrait de tous les marqueurs du chromosome 19 (geno-19) (4) le retrait de tous les marqueurs du chromosome 19 dans les génotypes 50k et l'ajout du pseudo SNP de groupe (geno-19_groupe). L'ensemble des scenarios présentés a été comparé à un ssGBLUP conduit sur les génotypes 50k à l'aide d'un test de Hotelling-Williams (Williams, 1959) implémenté sous R à l'aide du package *multilevel*.

Les précisions obtenues en single-step GBLUP pour l'ensemble des scenarios sont présentées sur la Figure 30.

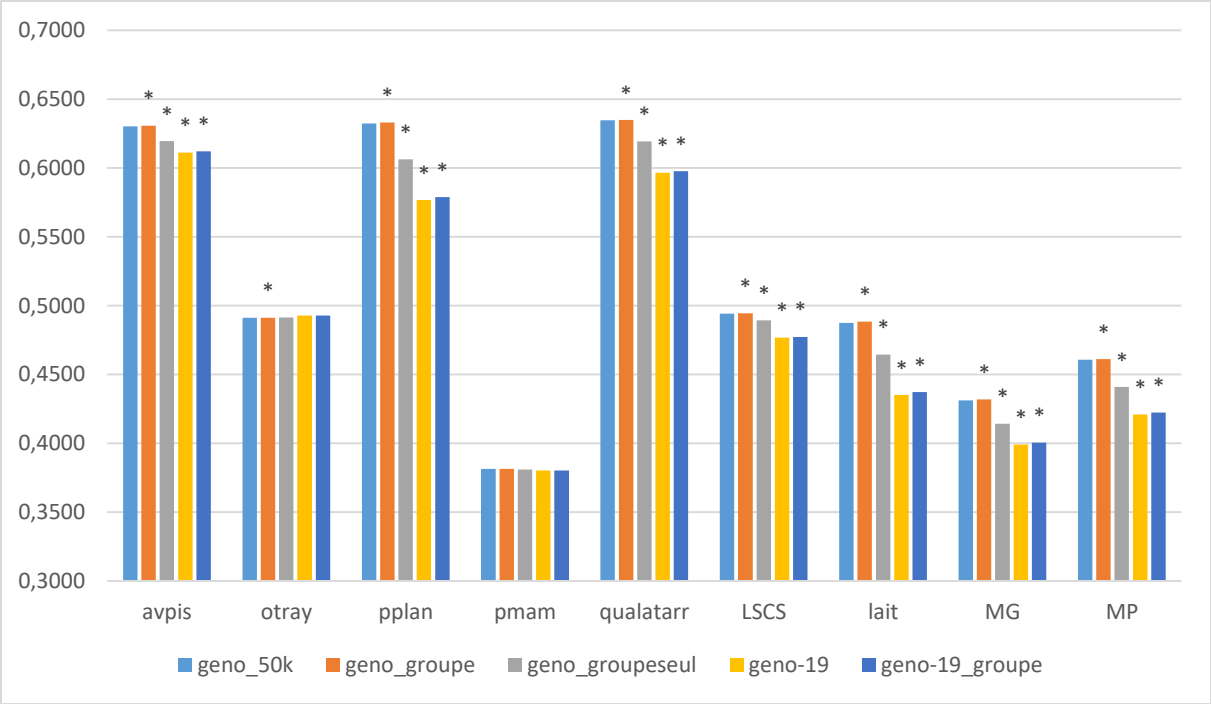


Figure 30: précision des évaluations sur les 148 Saanen de la population de validation obtenues suite à des ssGBLUP

* : significativement différents d'un ssGBLUP sur les génotypes 50k ($p < 0.05$)

abréviations : avpis : avant-pis ; otray : orientation des trayons ; pplan : position du plancher ; pmam : profil de la mamelle ; qualatarr : qualité de l'attache arrière ; MG : matière grasse, MP : matière protéique

L'ajout de l'information de groupe (geno_groupe) a conduit à de légères améliorations significatives de précision des évaluations génomiques. Toutefois ces dernières sont marginales : au maximum +0,2% de précision pour le lait et la MG. Le retrait de la région du QTL ou du chromosome 19 dans les génotypes 50k (scenarios geno_groupeseul, geno-19, geno-19_groupe) ont conduit à des dégradations significatives de la précision des évaluations génomiques pour tous les caractères à l'exception des caractères qui ne sont pas associés au chromosome 19 (otray et pmam). Ainsi, le scenario geno-19 a fait perdre en moyenne 6,9 points de précision aux évaluations de ces caractères. De plus, on note que l'information de groupe n'est pas suffisante pour remplacer l'information fournie par le chromosome 19 ou la région QTL qu'il porte. En moyenne, le scenario geno_groupeseul a conduit à une perte de 3,2% de précision pour les caractères lait, MG, MP, avpis, pplan, qualatarr et LSCS. Le scenario geno-19_groupe conduit, quant à lui, à une perte moyenne de précision de 6,6% pour les caractères associés au chromosome 19. On peut supposer que le pseudo-SNP avec son seul poids voit son information noyée dans l'ensemble des marqueurs de la puce. Il serait peut-être intéressant de lui attribuer artificiellement un poids particulier pour compenser ce phénomène.

On en conclut donc que l'information de groupe encodée sous la forme d'un SNP n'apporte que très peu d'information supplémentaire lorsque les marqueurs 50k de la région sont inclus dans les évaluations. Etant donnés les temps de calculs nécessaires à la construction du pseudo SNP, il paraîtrait plus judicieux de recourir à la version mise à jour de la puce (voir Article 4).

II. Estimation du biais des évaluations

En complément des analyses présentées dans l'article de ce chapitre (II.2), nous avons estimé le biais pour chacun des modèles et chacun des scenarios testés. Les Figures 31 et 32 présentent les résultats pour les scenarios ayant recours au génotypes 50k seuls, génotypes 50k avec les 178 variants de l'addon et les génotypes 50k complétés des 69 416 variants imputés de la région QTL. La Figure 31 présente les résultats obtenus en ssGBLUP et la Figure 32 en WssGBLUP.

En ssGBLUP (Figure 31), on observe que la densification de la région QTL en marqueurs a peu d'effet sur le biais pour les caractères qui ne sont pas associés au chromosome 19 (orientation des trayons et profil de la mamelle). Pour les caractères de productions laitières, les estimations de biais sont, quant à eux, plus proches de 1 suite à l'ajout de variants supplémentaire. L'amélioration des biais est maximale lorsque les évaluations sont faites sur génotypes 50kv2. Pour le reste des caractères à l'exception de la qualité de l'attache-arrière, le biais des prédictions s'éloigne de 1 avec la densification en marqueurs.

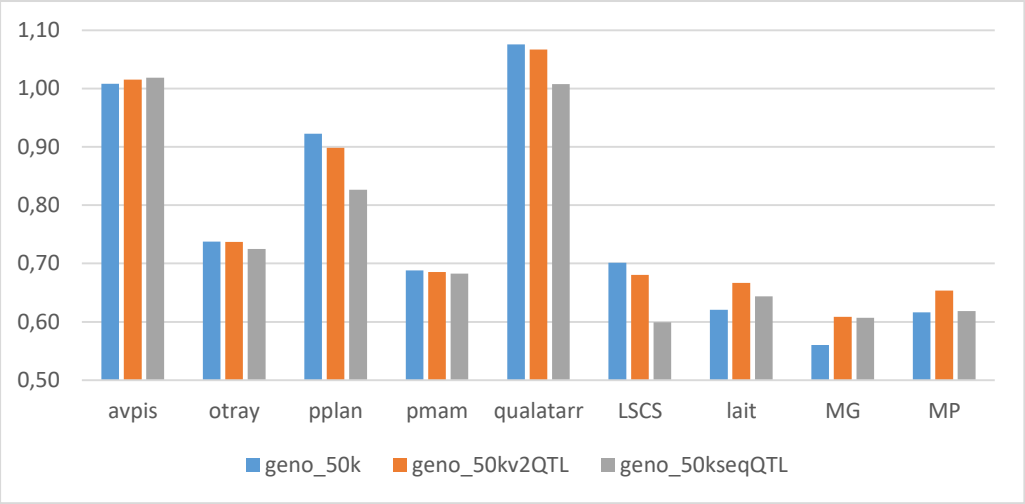


Figure 31: Biais (1 – coefficient de régression) observés pour des évaluations ssGBLUP sur genotypes 50k (geno_50k), genotypes 50 + 178 variants de l'addon (geno_50kv2QTL) et génotypes 50k complétés des variants de la région QTL du chromosome 19 (geno_50kseqQTL)

abréviations : avpis : avant-pis ; farrpis : forme de l'arrière pis ; ftray : forme des trayons ; itray : inclinaison des trayons ; ltray : longueur des trayons ; opied : ouverture des pieds ; otray : orientation des trayons ; pmam : profil de la mamelle ; pplan : position du plancher ; qualatarr : qualité de l'attache arrière ; tpoit : tour de poitrine

Les mêmes tendances sont observées avec un WssGBLUP (Figure 32). Elles sont cependant légèrement moins marquées. En revanche, nous notons une influence non-négligeable du modèle. En effet, les biais sont bien plus importants avec l'utilisation d'un WssGBLUP qu'un simple ssGBLUP. Ainsi, pour des évaluations sur les seuls génotypes 50k ; le biais moyen était de 0,77 en ssGBLUP et de 0,60 en WssGBLUP. Ceci a déjà été observé en caprins laitiers par Marc Teissier pendant ces travaux de thèse (Teissier, 2019). Chez d'autres espèces, le même phénomène a été observé. Ainsi en bovins laitiers, Lourenco et al. (2014) ont ainsi observé des biais plus proches de 1 en Holstein avec un ssGBLUP. De même, lors de l'évaluation d'une résistance à une maladie de la truite, Vallejo et al. (2017) ont estimé des biais de 0,86 en ssGBLUP et de 0,68 en WssGBLUP ce qui correspond aux variations que nous avons pu constater pour l'ensemble de nos caractères en caprins.

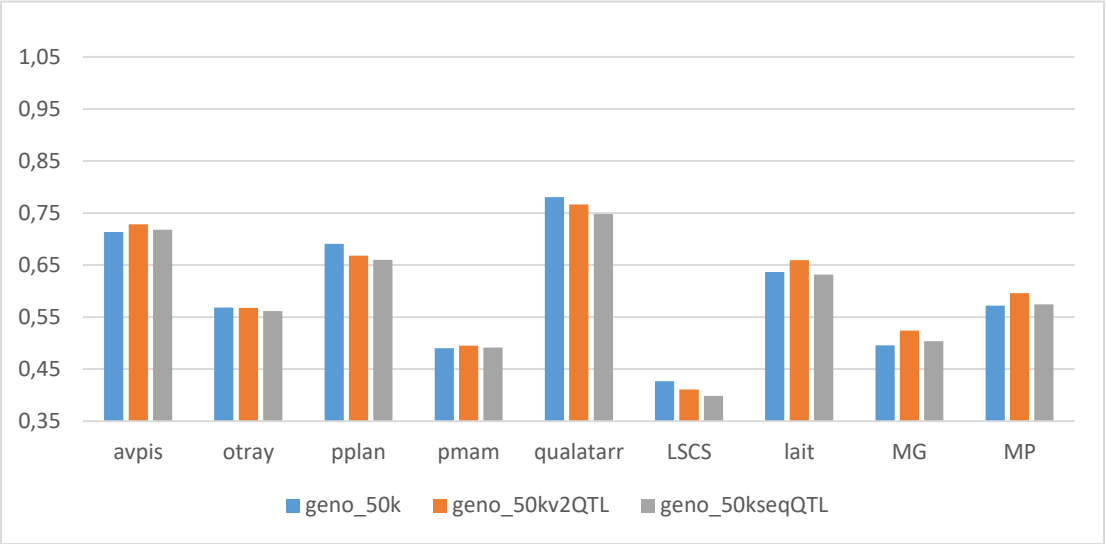


Figure 32: Biais (1 – coefficient de régression) observés pour des évaluations WssGBLUP sur genotypes 50k (geno_50k), genotypes 50 + 178 variants de l'addon (geno_50kv2QTL) et génotypes 50k complétés des variants de la région QTL du chromosome 19 (geno_50kseqQTL)

abréviations : avpis : avant-pis ; farrpis : forme de l'arrière pis ; ftray : forme des trayons ; itray : inclinaison des trayons ; ltray : longueur des trayons ; opied : ouverture des pieds ; otray : orientation des trayons ; pmam : profil de la mamelle ; pplan : position du plancher ; qualatarr : qualité de l'attache arrière ; tpoit : tour de poitrine

En conclusion, sur les biais des prédictions la puce 50kV2 semble là encore être une piste intéressante car elle améliore les précisions tout en limitant l'accroissement des biais liés à la densification en marqueurs. Les biais sont globalement améliorés pour l'ensemble des caractères de production et nous n'avons pas noté de dégradation majeure pour l'ensemble des autres caractères. Le WssGBLUP semble considérablement biaiser les évaluations et un compromis sera à trouver entre précision des évaluations et biais.

Conclusion du chapitre

Dans ce chapitre nous avons détaillé comment l'information issue des séquences et de leur analyse peut être utilisée dans les évaluations génomiques en Saanen française. Ces données peuvent être issues de l'imputation de séquence ou d'analyses des séquences (c'est le cas notamment de l'information de groupe en Saanen).

L'utilisation de données de séquences restreinte à des régions QTL particulières semble être une piste à envisager car elle permet d'améliorer significativement et de manière importante la précision des évaluations pour les caractères présentant des QTL connus. L'intégration de ces séquences ne semble pas significativement dégrader les évaluations des

caractères pour lesquels aucun QTL n'a pu être identifié. Ceci nous rend confiants quant à la possibilité d'intégrer la séquence dans les évaluations génomiques de routine. En effet, en routine, l'évaluation des caractères de production est conduite en multi-race, c'est-à-dire Alpine et Saanen conjointement. On peut alors espérer que l'ajout de la séquence de QTL identifiés en Saanen dans les évaluations multi-races ne dégrade pas la qualité des évaluations en Alpine. Cette assertion serait toutefois à vérifier avant toute implémentation.

L'utilisation de données de séquence en routine est toutefois source de contraintes liées au coût et au temps de préparation des données : établissement et entretien d'un panel de référence d'individus séquencés, imputation systématique des génotypes 50k vers la séquence, extraction des marqueurs des régions QTL dans les séquences imputées, préparation des fichiers pour les évaluations etc… La mise à jour de la puce apparaît dans ce contexte comme une alternative efficace et fiable. En effet, les génotypes obtenus sur les nouveaux typages seront plus sûrs que les génotypes imputés et, comme nous l'avons vu, permettraient de conduire des évaluations génomiques de qualité. Il faudra toutefois une étude spécifique pour évaluer si l'ensemble des marqueurs de l'addon peut être ajouté en routine aux évaluations. Il est, de plus, possible que tous les marqueurs sélectionnés dans la région du QTL (178 au total) ne passent pas l'étape de validation de l'addon avec un cluster file.

Chapitre 5 :
Discussion générale

Les résultats de ma thèse ayant été discutés au fil des différents chapitres précédents, nous nous attarderons, dans ce chapitre, à discuter des perspectives qui pourraient être envisagées suite aux travaux de cette thèse.

I. Amélioration de la fiabilité des imputations vers la séquence

En caprins laitiers français, comme évoqué dans le chapitre 1 paragraphe 6, le DL est peu persistant. En effet, deux marqueurs consécutifs de la puce 50k ne sont corrélés qu'à hauteur de 0,17 en moyenne (Carillier et al., 2013). Ceci complique le processus d'imputation. En effet, ce dernier repose sur le phasage des génotypes d'un individu. Lorsque le DL est faible, la certitude des phases diminue rapidement à mesure que l'on s'éloigne d'un marqueur de la puce. Dans notre étude, faute de puce haute-densité, nous prédisons des phases sur des distances moyennes de 60kb (espacement moyen de 2 marqueurs sur la puce 50k). L'incertitude des phases peut donc conduire à de multiples erreurs d'imputation ce qui peut expliquer que nos résultats ne soient pas aussi élevés que dans d'autres espèces, en particulier quand ces dernières ont pu bénéficier d'une imputation en plusieurs étapes à l'aide de puce à haute densité. Plusieurs pistes d'amélioration en caprins laitiers peuvent être envisagées : (1) séquencer plus d'individus pour obtenir une meilleure couverture de la diversité haplotypique des races Alpines et Saanen (2) encourager le développement d'une puce haute-densité (3) généraliser l'utilisation de la puce 50k développée cette année avec l'ajout de nouveaux marqueurs.

Concernant le premier point, il serait intéressant de garder à jour le panel en ajoutant régulièrement quelques individus récents. Cela permettrait de mieux suivre l'évolution des fréquences alléliques et de mieux évaluer les recombinaisons. De plus, la taille de la population de référence pour l'imputation est un facteur connu pour influencer la qualité des imputations (Frischknecht et al., 2017; Van Binsbergen et al., 2014; Ye et al., 2018). Ainsi, Van Binsbergen et al. (2014) ont augmenté la précision de l'imputation directe de la puce bovine 50k vers la séquence de 24% en utilisant 80% des 114 Holsteins séquencés au lieu de 40%. Cette étude ayant obtenu des corrélations similaires aux nôtres, si nous nous appuyons sur les résultats précédents, nous pouvons espérer atteindre une corrélation moyenne de 0.30 en Saanen et 0.32 en Alpine en séquençant 13 Saanen et 16 Alpines supplémentaires. Un

effort supplémentaire de séquençage dans les deux races serait donc à prévoir pour que les corrélations atteignent des valeurs correctes (supérieures à 0,90 par exemple). Enfin ces chiffres ne sont que des estimations, ils représentent un minimum d'individus supplémentaires à séquencer, en effet, les tailles de population efficace en caprins sont bien supérieures à celles observées en Holstein. En Saanen, la taille effective de population estimée est comprise entre 98 et 99 (travaux de thèse de Céline Carillier). En Alpine, elle a été estimée entre 115 et 127. Ainsi avec une centaine d'individus séquencés dans chaque race, nous pourrions espérer couvrir une grande partie de leur diversité.

Le développement d'une puce de génotypage haute-densité (HD) telle que disponible en bovins n'est, pour l'instant, pas envisagé par la filière caprine laitière française. D'après des prospectives récentes du consortium international génomique caprin (IGGC), les coûts de génotypages avec une telle puce HD seraient prohibitifs. La filière et l'IGGC ont donc préféré renégocier les prix de génotypages avec la puce 50k, encourager le développement de puces moyennes densité par la concurrence (ThermoFisher, Neogen…) et l'ajout de marqueurs à ces outils existants. Le recourt à des techniques comme le Genotyping by Sequencing (GbS) (Elshire et al. 2011) ou le RADseq (Restriction site-Associated DNA sequencing) (Miller, Dunham, Amores, Cresko, & Johnson, 2007) pourrait toutefois être une solution pour produire des génotypes équivalents en densité à une puce HD tout en limitant les coûts de génotypage. Les deux techniques sont équivalentes et reposent sur le séquençage ciblé de régions du génome suite à la digestion de l'ADN par des enzymes de restriction. Les fragments obtenus sont ensuite amplifiés par PCR puis séquencés produisant des lectures de quelques dizaines de paires de base. Le GbS a été décrit pour la première fois par Elshire et al. (2011). Cette technique est prometteuse car elle permet d'obtenir rapidement et de manière fiable les génotypes sur plusieurs sites du génome. En caprins, elle est d'autant plus prometteuse que 1 938 marqueurs correspondant à des SNPs identifiés par GbS (Rudiger Brauning, AgResearch) ont été ajoutés sur la version 2 de la puce 50k. Ces marqueurs en commun entre la puce v2 et la potentielle puce HD GbS pourraient faciliter l'imputation des génotypes 50k vers des génotypes HD. Cependant, le GbS présente une limite majeure : la quantité de génotypes manquants peut être élevée selon la profondeur de séquençage qui est définie. A titre d'exemple, lors d'une étude sur le Blé, Alipour et al. (2019) ont obtenu 133 039 SNPs en génotypant 384 individus, parmi ces derniers, 16 506, 38 642 et 65 560 avaient des pourcentages de génotypes manquant inférieurs à 20%, 50% et 80% respectivement. Il faudrait donc préalablement imputer les génotypes GbS disponibles pour compléter les génotypes manquants et obtenir un panel haute-densité consolidé et uniforme. En parallèle, de nouveaux outils d'imputation ont été développés (Zheng et al. 2018; Swarts et

al. 2014; Fragoso et al. 2016) et ont donné des résultats intéressants. Toutefois, pour assurer la qualité de cette première étape d'imputation, il faudra génotyper plusieurs centaines d'animaux dans chacune des races étudiées. Ainsi, Alipour et al. (2019), dans une étude avec 384 plants de Blé et sur les 65 560 SNPs initiaux qui avaient un *call rate* supérieur à 80%, obtiennent des taux de concordances compris entre 84,16% et 89,80% en fonction du génome de référence. Bien que l'imputation des génotypes issus de GbS ait déjà fait l'objet de plusieurs études, l'imputation de génotypes GbS vers la séquence reste encore à explorer. Cette piste est donc intéressante car elle n'a aucun précédent et pourrait s'avérer particulièrement adaptée aux petits ruminants car peu onéreuse. Le Tableau 9 résume les coûts des puces de génotypage en ovins. Ces prix sont purement indicatifs et sont issus d'un marché négocié en 2019 entre INRAE et Illumina. A ces derniers s'ajoutent des coûts d'extraction de l'ADN (4€) et de génotypage (11€). En caprins, les prix peuvent différer car les volumes sont inférieurs, les tendances peuvent cependant être similaires à celles observées en ovins. Pour le GbS, une équipe néozélandaise a développé deux pipelines dont les coûts sont largement inférieurs à ceux d'un génotypage classique (Shannon Clarke, AgResearch, communication personnelle). Le premier permet d'obtenir un set de variants à faible profondeur pour 25 NZD (soit environ 14€). Ce coût inclut l'extraction de l'ADN, la préparation des bibliothèques GbS et le séquençage. La technique permet d'obtenir entre 30k et 40k SNP sans génome de référence. La chèvre bénéficiant d'un génome de référence, le nombre de SNP identifiés par cette méthode peut être multiplié par 4 à 6. Ces derniers peuvent alors être utilisés dans les évaluations génomiques pour calculer une matrice de parenté génomique dont les coefficients tiennent compte de la profondeur de séquençage (Dodds et al., 2015). Le 2ème pipeline permet d'obtenir des variants qui pourront permettre de nouvelles analyses d'association. Ceux-ci sont séquencés avec une profondeur plus élevée ce qui rend les génotypes plus fiables. Le coût pour ce pipeline est compris entre 30 et 40 NZD (soit 17 à 23€) et permet d'obtenir entre 100k et 300k marqueurs. Les analyses d'association peuvent alors être conduite en utilisant des logiciels qui prennent en compte la probabilité de génotype. Ce pipeline pourrait également fournir des SNP pour une imputation par pallier en utilisant un logiciel comme Minimac capable d'utiliser des probabilités de génotype. Le coût reste, de plus, inférieur à celui d'un génotype 50k.

Tableau 9: Coûts de génotypage par puce en ovins en 2019

TYPE DE PUCE	*PRIX DE LA PUCE*	*NOMBRE DE MARQUEURS*
LD	16€	16 399
MD	13,70 à 16,48€	59 065
HD	90€	606 006

Abréviations : LD : Basse densité ; MD : Moyenne densité ; HD : Haute densité

Certaines des enzymes utilisées pour le GbS sont sensibles à la méthylation. La puce GbS pourrait ainsi permettre d'initier les premières analyses sur des variants de méthylation. Enfin la méthylation de l'ADN peut rendre son accès plus ou moins facile à l'ARN polymerase et donc influencer l'expression des gènes. Cette piste, intéressante, pourrait permettre d'identifier de nouvelles régions d'intérêt dans nos races d'études.

Comme évoqué précédemment, des marqueurs GbS ont été nouvellement ajoutés sur l'add-on de la puce 50k. Dans la filière française caprine, l'ensemble des nouveaux candidats à la sélection vont être génotypés avec cette puce mise à jour dès sa mise en service (horizon 2021). Ceci permettra d'obtenir 400 génotypages par an (Alpines et Saanen confondues) et ainsi d'intégrer potentiellement le GbS à la routine de la filière.

Avant de s'orienter vers l'utilisation de puces GbS, il est possible dans un futur plus proche d'utiliser la puce caprine Illumina moyenne densité dont le contenu augmenté en nombre de marqueur est en cours de validation. Les travaux de ma thèse ont permis de contribuer à cet add-on par la sélection de variants sur le chromosome 19. Les 178 marqueurs nouvellement identifiés, s'ils sont validés, pourraient grandement améliorer la qualité d'imputation de la région QTL. En effet, la région du génome comprise entre 23 et 30Mb sur le chromosome 19 présente des qualités d'imputation plus faibles que sur le reste du génome en race Saanen (Tableau 10). Enfin, l'add-on de la puce comporte 5 109 marqueurs. Parmi ces derniers, 414 ont été sélectionnés dans le but de densifier des zones avec un fort ou un faible taux de recombinaison (sélection effectuée par Rachel Rupp, GenPhySE, INRAE). Par ailleurs, 1 487 autres SNP ont été choisis dans des gammes de MAF faibles. L'ensemble de ces nouveaux variants pourraient permettre de mieux construire les haplotypes et de mieux estimer leur fréquence dans les différentes races. Ce nouvel outil pourrait donc permettre d'améliorer globalement la qualité de l'imputation, au-delà de la région du chromosome 19.

Afin de valoriser au mieux ces nouvelles données, des logiciels d'imputation différents pourront éventuellement être testés, une fois la carte génétique affinée (c'est-à-dire les hotspots de recombinaison identifiés). Certains logiciels de phasage et imputation peuvent, par exemple, prendre en compte la carte génétique ce qui permet d'améliorer la qualité des phases obtenues. C'est le cas, par exemple, de ShapeIt (Delaneau & Marchini, 2014) ou BEAGLE (Browning & Browning, 2011) par exemple.

Tableau 10:Qualité d'imputation de la puce 50kv1 vers la séquence évaluée sur une imputation uni-raciale en Saanen (33 individus)

	CORRELATION	***CONCORDANCE GENOTYPIQUE***	***CONCORDANCE ALLELIQUE***
CHI19 (QTL)	0,10	0,70	0,84
TOUT GENOME	0,24	0,74	0,86

II. Recherche de mutations causales

II.1. Retour sur les analyses d'association

Malgré la multiplicité des approches envisagées, notre étude approfondie des données de séquence du chromosome 19 en race Saanen n'a permis que d'affiner modestement la région du QTL précédemment identifiée sur les génotypes 50k. Cette incapacité à pointer un variant illustre les limites de l'analyse d'association : dans notre cas, le DL est important dans la région qui, par ailleurs, présente une très grande densité en gènes. D'autres approches s'appuyant sur des dispositifs familiaux de détections de QTL ont été envisagées dans le cadre de la thèse de Marie-Pierre Sanchez en bovins laitiers. Des analyses d'associations et analyses de liaisons sur des génotypes 50k ont constitué une première approche pour identifier des régions d'intérêt associées aux différentes protéines du lait dans les races d'étude : Holstein, Normande et Montbéliarde (Sanchez et al. 2016). Dans cette étude, les données de séquence ont été exploitées pour la recherche de mutations candidates qui ségrégeraient dans les familles de la même façon que le QTL identifié. Cette analyse a permis d'écarter des candidats (mutation Y581S du gène ABCG2 par exemple) tout en en confirmant d'autres (mutations du gène LGB ou la mutation F279Y du gène GHR). La filière caprine dispose également d'un dispositif familial dont les pères fondateurs ont été séquencés (9 Saanen et 11 Alpins). Toutefois, une étude similaire à celle effectuée par Marie-Pierre Sanchez qui s'est appuyée sur des qualités de génotype (GQ) supérieurs à 30, n'est pas envisageable en caprins

laitiers français. Les pères du dispositif de détection de QTL présentent, en effet, des séquences de piètre qualité. Ainsi, parmi les 20 individus séquencés du dispositif, 8 (4 Alpins et 4 Saanen) ont été écartés du jeu de séquence car leur profondeur moyenne de séquence était inférieure à 5 et leur qualité de génotype (GQ) était inférieure à 20.

Nous avons pu le constater au cours des travaux de cette thèse, le nombre de phénotypes disponibles et leur précision peut modifier le signal observé. Ainsi le passage de 490 DYD (boucs avec plus de 10 filles) en 2017 à 546 en 2020 a modifié le nombre de variants significatifs observés sur le chromosome 19 et la forme du signal dans la région QTL (Figure 33). Le signal central observé entre 26 et 27 Mb a légèrement perdu en significativité au profit de signaux périphériques. En particulier, des signaux supplémentaires sont apparus entre 28 et 29Mb avec l'ajout des nouveaux phénotypes. Néanmoins, l'acquisition de phénotypes et de méioses supplémentaires n'ont pas permis d'affiner le signal comme espéré.

En s'inspirant des travaux conduits dans la thèse de Marie-Pierre Sanchez chez les bovins laitiers, utiliser des prédictions fines de la composition du lait (en acides gras et protéines), pourraient permettre d'obtenir des phénotypes plus précis qui pourraient être associés à une région plus restreinte du chromosome 19, et également permettre l'identification de nouvelles régions QTL sur l'ensemble des chromosomes. Marie-Pierre Sanchez a utilisé des phénotypes très fins pour les protéines du lait grâce à l'analyse de spectre MIR. Des analyses d'association sur séquences imputées de plus de 8 000 vaches ont permis d'affiner les régions QTL notamment en procédant à des analyses conditionnelles. Le variant candidat est alors ajouté en effet fixe dans un modèle d'analyse d'association classique. En sortie, si plus aucun des autres variants n'est significatif dans la région, c'est que ce candidat capte la majeure partie de la variance expliquée par la région QTL. Il peut alors être considéré comme un excellent candidat et faire l'objet d'analyses fonctionnelles. Suite à l'identification de plusieurs variants candidats annotés pour plusieurs gènes, Marie-Pierre Sanchez a recherché des voies métaboliques associées à ces derniers. Ses analyses ont alors permis d'identifier des liens entre les gènes candidats (Sanchez et al. 2019). Le lien de causalité exact entre ces gènes et les variations du phénotype restent toutefois à démontrer. En caprins laitiers, des projets se sont déjà intéressés à la composition fine du lait (en acides gras) cependant l'analyse des spectres MIR n'a pas permis d'établir des équations de prédiction des protéines suffisamment précises pour être exploitées dans des analyses d'association et espérer cartographier finement un QTL.

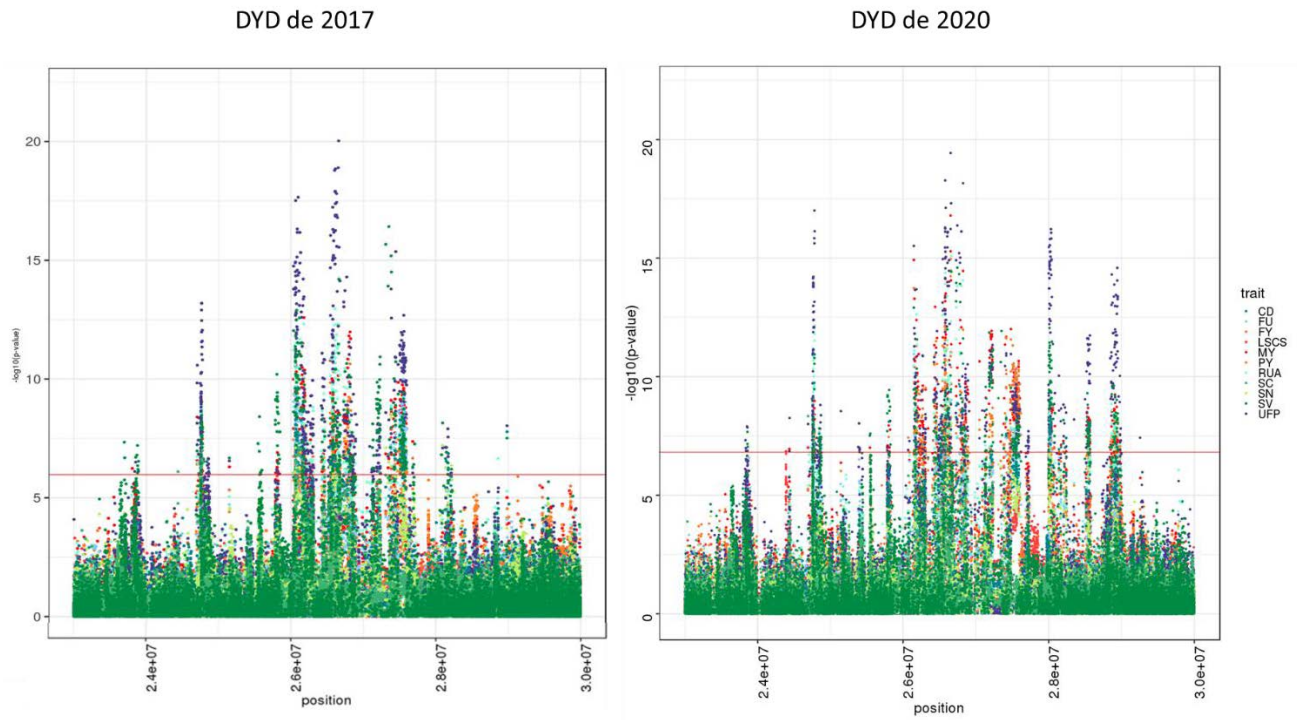


Figure 33: Manhattan plots des analyses d'association effectuées en Saanen sur 490 mâles (2017) et 546 mâles (2020)

CD : tour de poitrine ; FU : avant-pis ; FY : matière grasse ; LSCS : cellules somatiques du lait ; MY : lait ; PY : matière protéique ; RUA : qualité de l'attache-arrière ; SC : concentration en spermatozoïdes ; SN : nombre de spermatozoïdes ; SV : volume de semence ; UFP : position du plancher

II.2. L'exhaustivité des données de séquences

Les données de séquence, par leur exhaustivité, contiennent *a priori* les variants causaux et portent en elles l'espoir d'identifier la mutation causale d'un phénotype donné. En pratique, cette dernière n'est pas toujours présente dans le jeu de données. En effet, il faut que celui-ci comporte suffisamment d'individus porteurs des différents allèles, que l'assemblage local soit de bonne qualité et que les séquences individuelles soient suffisamment profondes à l'endroit de la mutation pour assurer un génotype fiable. Par exemple, sur les 5 mutations faux-sens qui permettent de discriminer les profils de caséines en caprins, seules 4 ont passé le filtrage que nous avons imposé. En effet, la qualité des variants de la région est relativement faible dû à une faible qualité d'assemblage de la séquence de référence dans la région. La mutation qui induit la substitution d'une histidine par une isoleucine a été éliminée. Deux raisons possibles à cette élimination : la mutation était trop proche d'un autre variant de meilleure qualité ou elle ne respectait pas la condition suivante : 95% des animaux avec GQ

>=7 ou DP >=8. Toutefois, à supposer que la mutation causale soit présente dans le jeu de données, son identification n'en reste pas moins complexe.

II.3. Comment prioriser les variants pour une étude fonctionnelle ?

Nous avons tenté d'exploiter au maximum l'information disponible et connue à ce jour pour les variants significatifs dans notre région d'intérêt. Etant donnée la largeur du signal observé sur le chromosome 19, tous les variants ne peuvent pas faire l'objet d'une analyse fonctionnelle approfondie. Dans ce cas de figure, prioriser les variants devient essentiel.

La recherche dans l'annotation de variants avec un fort impact sur la structure d'une protéine ou sur l'expression d'un gène peut être une piste intéressante. L'identification d'éléments fonctionnels permettra de mieux comprendre le lien entre génotype et phénotype. En effet, dans une région QTL telle que celle que nous avons étudiée sur le chromosome 19, cette approche pourrait permettre de distinguer les variants qui ont un impact, avéré ou supposé, de ceux qui sont simplement en déséquilibre de liaison avec la mutation causale. A l'heure actuelle, la qualité de l'annotation du génome caprin est encore insuffisante pour que cette approche soit pleinement opérationnelle, mais elle fait l'objet d'un groupe de travail dans le projet VarGoats.

En caprins, et en particulier en race Saanen française, il serait intéressant de procéder à un séquençage de l'exome (partie exprimée du génome) sur la région du QTL du chromosome 19. Nous aurions alors accès à l'ensemble des ARN exprimés dans les différents tissus en lien avec les caractères associés au CHI19 : tissus mammaires, testicules, cellules du système immunitaire par exemple et pourrions alors effectuer une analyse de QTL d'expression. La région du QTL est très riche en gènes : 23,4 gènes par Mb contre 10,6 gènes en moyenne sur l'ensemble des autosomes (https://www.ncbi.nlm.nih.gov/genome/?term=Capra_hircus; consulté le 23/04/2020). Une approche par QTL d'expression réduirait le nombre de gènes à investiguer et pourrait parallèlement considérablement réduire le nombre de variants à étudier pour les analyses fonctionnelles. Néanmoins, la mutation causale peut ne pas affecter l'expression des gènes ou peut affecter l'expression de gènes ailleurs sur le génome.

Il nous paraît tout de même important de ne pas exclure les variants qui ne se trouvent pas dans les exons d'un gène. Nous avons, en effet, pu identifier des variants dans des micro-ARN (miARN) sur le chromosome 19 : MIR195, MIR324 et MIR497. Les miARN sont de petits brins d'ARN de quelques paires de bases qui ne se traduisent pas en protéine mais régulent la traduction d'ARN porteurs de séquences complémentaires. Pour compléter nos

analyses, il serait alors intéressant d'identifier les cibles de ces miARN. Chez l'Homme, il a été démontré que les miARN peuvent avoir plusieurs centaines de cibles différentes et impacter le profil d'expression général d'un tissu (Lim et al., 2005). Ils constituent une cible intéressante dans le cas de notre région pléiotrope car la possible multiplicité des cibles des miARN pourrait expliquer que des caractères aussi variés que la production de lait, la production de semence et la conformation de l'animal soient associés à une même région du génome.

II.4. Recherche de variants structuraux

L'analyse des données brutes de séquence à l'aide du logiciel IGV est une approche manuelle fastidieuse. Elle nous a permis d'identifier un variant structural de grande taille. Cette approche doit être complétée par un génotypage local ou par vérification avec un *calling* des variants structuraux pour confirmer qu'il ne s'agit pas d'un artefact et pour correctement identifier les individus porteurs hétérozygotes. Si le génotypage n'est pas envisageable, nous avons vu qu'il était possible d'imputer correctement un génotype pour un variant structural à partir d'un nombre limité d'individus séquencés. Si le variant est biallélique, les méthodes classiques d'imputation pourront être appliqués facilement. L'imputation de variants structuraux a été étudiée en bovins laitiers avec des résultats corrects : R^2 compris entre 0,84 et 0,93 en fonction des logiciels utilisés pour le phasage (Mesbah-Uddin, Guldbrandtsen, Lund, Boichard, & Sahana, 2019). Dans le cas de variants multi-alléliques, il sera nécessaire, par exemple, de recourir à une conversion pour coder les allèles en 0, 1, 2. Ceci peut être fait en appliquant la même méthode que Marc Teissier a utilisée pour utiliser des haplotypes dans les évaluations génomiques (M. Teissier, 2019). Il suffit de référencer le nombre d'allèles puis de regarder lesquels sont portés par chaque individu (Tableau 11).

Tableau 11: Exemples de conversion du génotype pour un variant avec 4 allèles

GENOTYPE DE L'INDIVIDU	A_1	A_2	A_3	A_4	***GENOTYPE A FOURNIR AU LOGICIEL D'IMPUTATION***
A_1A_1	2	0	0	0	2000
A_2A_4	0	1	0	1	0101
A_1A_3	1	0	1	0	1010

II.5. Utilisation de données génomiques d'autres races caprines laitières

Le QTL identifié sur le chromosome 19 a été détecté en Saanen française mais est absent en race Alpine. Par ailleurs la validation menée au chapitre 3 suggère que le QTL

ségrége dans d'autres populations de chèvres Saanen dans le monde (article 3). Aujourd'hui, des échanges existent entre la France, la Suisse, l'Italie et le Canada dans le cadre du projet H2020 Smarter. Ces pays sont connectés par des échanges ponctuels d'animaux et/ou de semence (travaux de Marc Teissier en cours). Nous pourrions ainsi tirer profit de l'acquisition de séquences supplémentaires de Saanen. Le DL étant fort au sein de la population française, l'ajout d'animaux de populations extérieures porteuses du même QTL pourrait permettre de casser ce DL. Il sera, de plus, pertinent de conduire des analyses d'association sur ces dernières. Une analyse conjointe des races ou une méta-analyse des analyses d'association intra-races pourrait augmenter la puissance des dispositifs de détection et peut-être affiner la région du QTL du chromosome 19. Le coût de séquençage restant tout de même élevé, une alternative intéressante pourrait être l'utilisation massive de la puce caprine mise à jour pour acquérir des génotypes en Saanen à l'international. Il faudra également disposer de phénotypes précis pour chacun des animaux génotypés ou séquencés. De tels dispositifs ont été mis en place dans le cadre du projet européen SMARTER, débuté en 2018, (https://www.smarterproject.eu) pour explorer la faisabilité d'évaluations génomiques internationales en caprins.

Mucha et al. (2017) ont pu génotyper des femelles d'une race caprine synthétique britannique : 2 381 femelles avec des performances de production et 402 femelles pointées pour des caractères de morphologie. Suite à des analyses d'association, ils ont eux aussi trouvé un important QTL sur le chromosome 19. Ce dernier est situé dans la même région que notre QTL. Il est associé à la quantité de lait ainsi qu'à des caractères de conformation : les aplombs avant, la profondeur de la mamelle et l'attache-arrière. Cette race est issue d'un croisement qui a contenu de la Saanen. Ce croisement a été entretenu pendant plusieurs générations et la race est maintenant considérée comme stabilisée. Cette race est donc très intéressante car les crossing-over qui auraient pu avoir lieu depuis sa création pourraient nous aider à affiner la région du QTL en Saanen.

Enfin une approche utilisant les signatures de sélection est envisageable. Nous l'avons constaté, la région QTL sur le chromosome 19 est vaste et en fort déséquilibre de liaison. Ceci nous porte à croire que la sélection de cette région est récente. Nous ignorons toutefois comment celle-ci est apparue et comment elle a été sélectionnée. Dans un futur proche, une étude intra-population, c'est-à-dire intra-Saanen française, permettrait de comparer les fréquences de la région QTL du chromosome 19 à celles du reste du génome sous sélection neutre (Simonsen et al. 1995). Nous pourrions alors retracer l'historique de sélection de la

région. Cette étude pourrait être puissante sur la puce 50kv2 enrichie en marqueurs dans la région du QTL.

III. Intérêt des données de séquence dans les évaluations génomiques

III.1. Une puce 50k v2 prometteuse

La validation des SNP de la mise à jour de la puce 50k est intervenue trop tard pour que les nouveaux génotypes puissent faire l'objet d'une étude dans le cadre de ma thèse. Mon travail s'est, par conséquent, appuyé uniquement sur les SNP que j'avais choisis d'intégrer sur cet add-on. Ces SNPs sont concentrés dans la région comprise entre 23 et 30 Mb du chromosome 19. Ils étaient donc particulièrement pertinents pour les prédictions en race Saanen. Ainsi, les gains de précision moyens par rapport à la puce 50kv1 étaient compris entre 3,1 et 6,4% en fonction du scenario. L'ajout de ces marqueurs a notamment été bénéfique pour l'évaluation des caractères de production associés au CHI19 (lait, MG, MP). Il sera intéressant une fois que suffisamment de génotypes seront acquis (ou éventuellement imputés) de conduire de nouvelles analyses en Saanen et de lancer les premières évaluations en Alpine.

III.2. Vers une utilisation de la séquence de l'ensemble des régions QTL

Dans nos analyses, nous nous sommes concentrés sur la région QTL du chromosome 19. Toutefois, d'autres régions QTL existent en Saanen, la région des caséines (CHI6) (Martin and Leroux 2000) ou encore celle de DGAT1 (CHI14) (Martin et al., 2017). L'intégration du génotype des caséines a déjà été testée précédemment dans les travaux de thèse de Céline Carillier (2015) et de Marc Teissier (2019). Pour compléter l'approche que nous avons adoptée sur séquence, il serait intéressant d'étudier l'intégration des données de séquence des différentes régions QTL identifiées via les analyses d'association sur l'ensemble du génome dans les évaluations génomiques. Lorsque le QTL présente un fort effet sur le caractère alors le meilleur modèle semble être un WssGBLUP. Ce dernier a grandement amélioré la précision des prédictions pour les caractères laitiers par exemple (+15% en moyenne par rapport à un ssGBLUP sur génotypes 50k). Le WssGBLUP est à préférer au WssGBLUP avec des fenêtres ($WssGBLUP_{fenêtres}$). En effet, le $WssGBLUP_{fenêtres}$ est, par définition, légèrement plus exigeant en temps de calcul. Et les gains moyens de précision observés étaient minimes : compris entre 2 et 4% en fonction du typage utilisés (50k, 50kv2, séquence du QTL) et de la méthode de construction des fenêtres (2,4Mb ou 40 SNP consécutifs).

Toutefois, le WssGBLUP semble dégrader plus fortement la précision des prédictions des caractères qui ne présentent pas de QTL. Ainsi pour les cellules somatiques du lait, l'instauration d'un WssGBLUP a dégradé la précision des évaluations de 14,6% par rapport à un ssGBLUP. Ceci soulève plusieurs questions. Sera-t-il nécessaire d'établir un modèle par caractère et par race ? Ceci peut s'avérer extrêmement contraignant et en particulier dans le cadre des évaluations génomiques en routine. Il serait alors nécessaire de construire des modèles différenciés pour ces caractères. On aurait, par conséquent, un modèle pour le TB (incluant la région DGAT1 et la région des caséines en Saanen et incluant en plus une région QTL sur le chromosome 6 en Alpine), un modèle pour le TP (incluant uniquement la région des caséines en Alpine et Saanen) et enfin un modèle pour le lait et les matières (incluant la région du chromosome 19 en Saanen). Les évaluations actuelles devraient alors être modifiées puisqu'actuellement les caractères de production sont évalués à l'aide d'un modèle multi-race. Des études complémentaires seraient nécessaires afin de donner tous les éléments (avantages, inconvénients, limites) aux personnes en charge des évaluations et au partenaire professionnel (Capgènes) pour décider du ou des modèles futurs à adopter. En particulier, une étude préalable devrait être menée pour quantifier l'impact de l'intégration des variants de la région QTL du chromosome 19 en Saanen sur les évaluations du lait et des matières en Alpine. Nous avons vu que la puce 50kv2 était prometteuse en Saanen, cela reste à confirmer. Cette étude peut être menée en parallèle en Alpine et Saanen bien que les objectifs ne soient pas exactement les mêmes dans les 2 races (Figure 34). Elle devra alors explorer pour les deux

races, les coûts et bénéfices des différents modèles (ssGBLUP, WssGBLUP, $WssGBLUP_{fenêtres}$), du potentiel passage à une évaluation uni-raciale pour tous les caractères et enfin de l'adoption de modèles différents en fonction du caractère étudié.

III.3. Utilisation de l'annotation du génome

Les évaluations génomiques peuvent intégrer des informations supplémentaires à mesure que la connaissance des génomes animaux est approfondie. Ce lien est précieux pour mettre en place des évaluations génomiques toujours plus précises. Des projets tels que FR-AgENCODE vont apporter différents types de données qu'il serait intéressant d'intégrer aux évaluations. Ainsi, nous aurons une idée de quelle région du génome est exprimée (via des analyses RNAseq), quelles régions du génome sont accessibles aux protéines de la transcription (via de l'ATAC-seq) et enfin quelles régions du génome sont spatialement proches dans le noyau et peuvent donc potentiellement s'influencer. Il faudra alors développer de nouveaux modèles pour intégrer ces différentes informations.

Enfin, nous l'avons vu, il est théoriquement possible de convertir les variants

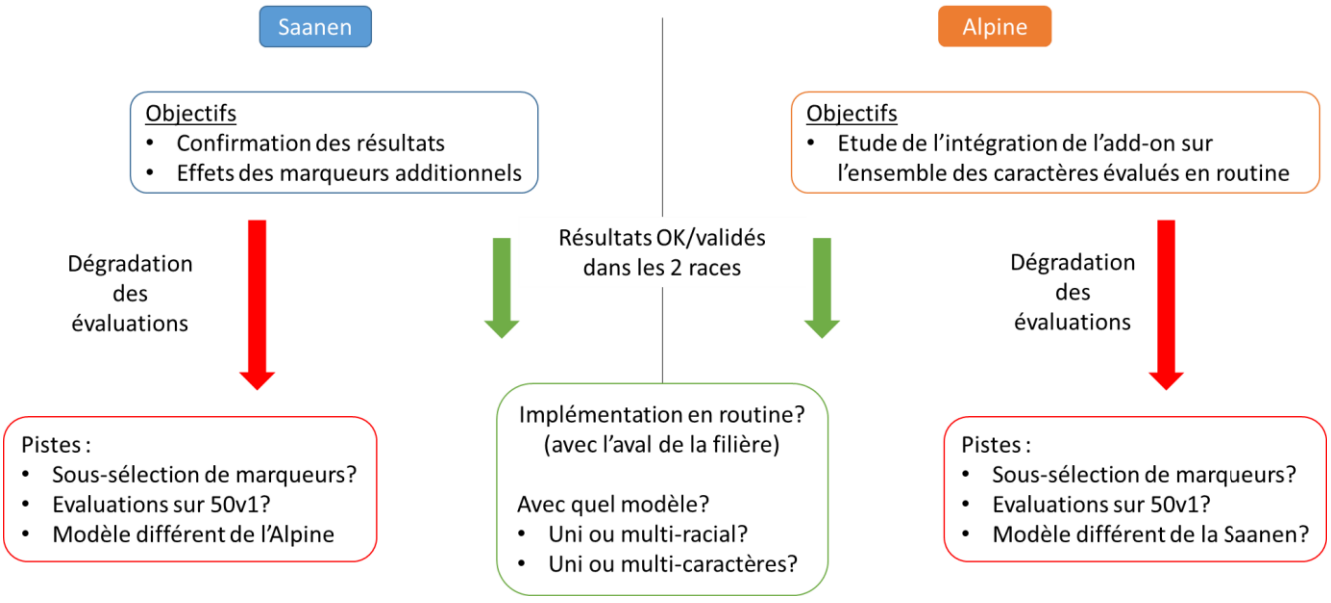


Figure 34: Plan d'étude de l'intégration des génotypes 50kv2 dans les évaluations génomiques de routine en race Alpine et Saanen

structuraux pour qu'ils puissent être intégrés dans les évaluations au même titre que des SNP

bi-alléliques classiques. Il sera alors intéressant de les intégrer dans les modèles d'évaluations génomiques et étudier les gains de précision.

III.4. Evaluations internationales

En bovins laitiers, des évaluations génétiques internationales ont été mises en place et ont permis des gains substantiels de précision par rapport à des évaluations intra-pays (Paul M Vanraden & Sullivan, 2010). Elles nécessitent cependant des infrastructures pour assurer la coopération et centraliser les capacités de calculs nécessaires à mettre en place des évaluations d'une telle ampleur. Le modèle utilisé est appelé MACE (Multiple-trait Across Country). Il s'appuie sur les EBV déregréssés obtenus dans les évaluations nationales des pays impliqués. Le pays de l'évaluation, ainsi que le groupe d'origine de chaque taureau sont également pris en compte. Les EBV sont ensuite fournis à chaque pays en tenant compte de l'adaptation aux conditions d'élevage nationales. Chaque pays est libre de sélectionner les animaux en fonction de son propre objectif de sélection.

Les races Alpine et Saanen sont présentes sur les 5 continents. L'un des objectifs du projet H2020 Smarter est de renforcer la coopération internationale autour de ces races afin d'améliorer les gains génétiques globaux autour des caractères de résilience et d'efficacité. Cette coopération internationale se traduit notamment par la mise en place d'évaluations internationales. Ainsi, la France, le Canada, l'Italie et la Suisse mettent en commun des données tels que les pedigrees, les phénotypes et les génotypes 50k de leurs populations. La connexion génétique des pays, les moyennes et variances des caractères, leurs héritabilités et les corrélations génétiques entre caractères pourront être évaluées. Et différents scenarios d'évaluations génomiques seront testés.

En Saanen, d'autres populations semblent porter un QTL sur le chromosome 19 (cf. Chapitre 3). Ce projet est donc une opportunité sans précédent d'intégrer des données issues de ces dernières et de mieux estimer l'effet de cette région et ainsi éventuellement améliorer les prédictions des évaluations (en particulier si un WssGBLUP est envisagé). En Alpine, il semble que les caractères d'intérêt mesurés actuellement, à l'exception du TB et du TP pour lesquels des mutations ont été identifiées, aient plutôt un déterminisme polygénique. Les prédictions dans cette race pourraient tout de même être améliorée par le simple apport d'individus.

Conclusion

Les données de séquence se sont largement démocratisées au cours des deux dernières décennies. Elles sont porteuses de nombreux espoirs dont certains ont été au cœur de mes travaux de thèse. Mon objectif était multiple (i) évaluer la faisabilité d'une imputation directe vers la séquence des génotypes disponibles (ii) utiliser les séquences imputées pour la détection de nouveaux QTL et la confirmation de QTL précédemment identifiés sur les génotypes 50k (iii) étudier l'intérêt d'introduire une partie des données de séquence dans les évaluations génomiques caprines.

Nous avons montré que l'imputation directe des génotypes 50k vers la séquence donnait des résultats corrects : les concordances alléliques et génotypiques ont, en effet, atteint respectivement 0,86 et 0,75 en Alpine et 0,86 et 0,73 en Saanen. Les corrélations étaient en moyenne de 0,26 et 0,24 respectivement dans ces deux races. L'imputation est moins précise que dans d'autres espèces d'élevage comme les bovins et les ovins. Elle permet toutefois de détecter correctement des régions QTL d'intérêt pour la filière caprine.

Nous avons mis à profit la disponibilité des séquences en Saanen pour approfondir le chromosome 19 porteur d'une région QTL d'environ 5 Mb associée à de nombreux caractères d'intérêt pour la filière laitière : caractère de production, conformation et stature de l'animal et enfin capacité à produire de la semence. Nous avons été en mesure de définir différents profils génomiques de Saanen qui correspondent à des profils phénotypiques distincts. Ces derniers peuvent être différenciés à partir de génotypages 50k. Cet outil pourra être utilisé par la filière pour servir son objectif de sélection. Il a également permis de confirmer la présence de ce QTL dans d'autres populations à l'échelle mondiale.

Enfin nous avons testé plusieurs stratégies d'intégration des données de séquence du chromosome 19 dans les évaluations génomiques en race Saanen. La mise à jour de la puce caprine actuelle semble être une approche prometteuse. L'intégration de 178 marqueurs choisis dans la région du QTL permet d'améliorer significativement la qualité des évaluations pour les caractères associés au QTL sans pour autant dégrader la qualité des évaluations des autres caractères.

L'ensemble des résultats produits au cours de cette thèse indique que les séquences sont intéressantes pour compléter les premières approches faites précédemment à l'aide de la puce de génotypage 50k. Toutefois, d'autres pistes doivent être envisagées pour améliorer la

qualité d'imputation vers la séquence et les détections de QTL qui s'en suivent. Parmi ces pistes : séquencer plus d'individus Alpine et Saanen, y compris des individus étrangers, utiliser massivement la puce mise à jour en France comme ailleurs et enfin potentiellement recourir à la technique du GbS pour obtenir une puce pseudo-HD.

Bibliographie

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). Hot topic : A unified approach to utilize phenotypic , full pedigree , and genomic information for genetic evaluation of Holstein final score 1. *Journal of Dairy Science*, *93*(2), 743–752. https://doi.org/10.3168/jds.2009-2730

Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, *27*(11), 2534–2547. https://doi.org/10.1093/molbev/msq148

Alipour, H., Bai, G., Zhang, G., Bihamta, M. R., Mohammadi, V., & Peyghambari, S. A. (2019). Imputation accuracy of wheat genotyping-by-sequencing (GBS) data using barley and wheat genome references. *PLoS ONE*, *14*(1), 1–20. https://doi.org/10.1371/journal.pone.0208614

Antolín, R., Nettelblad, C., Gorjanc, G., Money, D., & Hickey, J. M. (2017). A hybrid method for the imputation of genomic data in livestock populations. *Genetics Selection Evolution*, 1–17. https://doi.org/10.1186/s12711-017-0300-y

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Babo, D. (2000). *Races ovines et caprines françaises* (Editions F).

Barillet, F., Mariat, D., Amigues, Y., Faugeras, R., Caillat, H., Moazami-Goudarzi, K., … Perrin-Chauvineau, C. (2009). Identification of seven haplotypes of the caprine PrP gene at codons 127, 142, 154, 211, 222 and 240 in French Alpine and Saanen breeds and their association with classical scrapie. *Journal of General Virology*, *90*(3), 769–776. https://doi.org/10.1099/vir.0.006114-0

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview : analysis and visualization of LD and haplotype maps, *21*(2), 263–265. https://doi.org/10.1093/bioinformatics/bth457

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., … Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *49*(4), 643–650. https://doi.org/10.1038/ng.3802

Bindon, B. M. (1984). Reproductive Biology of the Booroola Merino Sheep. *Australian Journal of Biological Sciences*, *37*(3), 163–190. Retrieved from https://doi.org/10.1071/BI9840163

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., … Wiggans, G. R. (2012). Design of a bovine low-density snp array optimized for imputation. *PLoS ONE*, *7*(3), 1–10. https://doi.org/10.1371/journal.pone.0034130

Bolormaa, S., Chamberlain, A. J., Khansefid, M., Stothard, P., Swan, A. A., Mason, B., … MacLeod, I. M. (2019). Accuracy of imputation to whole-genome sequence in sheep. *Genetics Selection Evolution*, *51*(1), 1. https://doi.org/10.1186/s12711-018-0443-5

Boussaha, M., Michot, P., Letaief, R., Hozé, C., Fritz, S., Grohs, C., … Boichard, D. (2016). Construction of a large collection of small genome variations in French dairy and beef breeds using whole-genome sequences. *Genetics Selection Evolution*, *48*(1), 1–10. https://doi.org/10.1186/s12711-016-0268-z

Bouwman, A. C., & Veerkamp, R. F. (2014). Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. *BMC Genetics*, *15*(1), 1–9. https://doi.org/10.1186/s12863-014-0105-8

Brøndum, R. F., Guldbrandtsen, B., Sahana, G., Lund, M. S., & Su, G. (2014). Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics*, *15*(1), 1–8. https://doi.org/10.1186/1471-2164-15-728

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction, *8*, 4107–4116. https://doi.org/10.3168/jds.2014-9005

Browning, S. R., & Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering, *81*(November), 1084–1097. https://doi.org/10.1086/521987

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, *12*(10), 703–714. https://doi.org/10.1038/nrg3054

Carillier-Jacquin, C., Larroque, H., & Robert-Granié, C. (2016). Including α s1 casein gene information in genomic evaluations of French dairy goats. *Genetics Selection Evolution*, *48*(1), 1–13. https://doi.org/10.1186/s12711-016-0233-x

Carillier, C., Larroque, H., Palhière, I., Clément, V., & Rupp, R. (2013). A first step toward genomic selection in the multi-breed French dairy goat population. *Journal of Dairy Science*, *96*(11), 7294–7305. https://doi.org/10.3168/jds.2013-6789

Carillier, C., Larroque, H., & Robert-Granié, C. (2017). La sélection génomique chez les caprins laitiers français. *Inra Prod. Anim.*, *30(1)*, 19–30.

Cerda, M., Haitina, T., Schio, H. B., & Peter, R. E. (2005). Gene Structure of the Goldfish Agouti-Signaling Protein : A Putative Role in the Dorsal-Ventral Pigment Pattern of

Fish, *146*(3), 1597–1610. https://doi.org/10.1210/en.2004-1346

Chen, W., Kong, J., Qin, C., Yu, S., Tan, J., Chen, Y. R., … Hong, Y. (2015). Requirement of CHROMOMETHYLASE3 for somatic inheritance of the spontaneous tomato epimutation Colourless non-ripening. *Scientific Reports*, *5*, 1–7. https://doi.org/10.1038/srep09192

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics*, *3*.

Clément, V., Ceglowski, C., de Cremoux, R., Martin, P., & Rupp, R. (2015). Concentrations cellulaires dans les élevages caprins: états des lieux et mise en place d'un programme d'amélioration génétique. In *Renc Rech Rumin* (p. 22:45-48).

Clément, V., Martin, P., & Barillet, F. (2006). Elaboration d ' un index synthétique caprin combinant les caractères laitiers et des caractères de morphologie mammaire Elaboration of a total merit index combining dairy and udder type traits. *Renc. Rech. Ruminants*, (1), 209–212.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores , and the Solexa / Illumina FASTQ variants, *38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

Colleau, J.-J., Moureaux, S., Briend, M., & Bechu, J. (2004). A method for the dynamic management of genetic variability in dairy cattle. *Genetics Selection Evolution*, *36*(4), 373. https://doi.org/10.1186/1297-9686-36-4-373

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R. F., … Hayes, B. J. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, *46*(8), 858–865. https://doi.org/10.1038/ng.3034

Dassonneville, R., Fritz, S., Ducrocq, V., & Boichard, D. (2012). Short communication : Imputation performances of 3 low-density marker panels in beef and dairy cattle. *Journal of Dairy Science*, *95*(7), 4136–4140. https://doi.org/10.3168/jds.2011-5133

Delaneau, O., & Marchini, J. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, 1–9. https://doi.org/10.1038/ncomms4934

Dodds, K. G., Mcewan, J. C., Brauning, R., Anderson, R. M., Stijn, T. C. Van, Kristjánsson, T., & Clarke, S. M. (2015). Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, 1–15. https://doi.org/10.1186/s12864-015-2252-3

Drouilhet, L., Lecerf, F., Bodin, L., Fabre, S., & Mulsant, P. (2009). Fine mapping of the FecL locus influencing prolificacy in Lacaune sheep. *Animal Genetics*, *40*(6), 804–812.

https://doi.org/10.1111/j.1365-2052.2009.01919.x

Druet, T., & Georges, M. (2010). A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. https://doi.org/10.1534/genetics.109.108431

Druet, T., Macleod, I. M., & Hayes, B. J. (2014). Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity*, *112*(1), 39–47. https://doi.org/10.1038/hdy.2013.13

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), 1–10. https://doi.org/10.1371/journal.pone.0019379

Evol, S., Stella, A., Nicolazzi, E. L., Tassell, C. P. Van, Rothschild, M. F., Colli, L., … Joost, S. (2018). AdaptMap : exploring goat diversity and adaptation, 1–7. https://doi.org/10.1186/s12711-018-0427-5

Ewing, B., Ladeana, H., Michael, C. W., & Green, P. (1998). Base-Calling of Automated Sequencer Traces using Phred. I. Accuracy Assessment, (8), 1–11.

Ewing, B., LaDeana, H., Michael, C. W., & Green, P. (1998). Base-Calling of Automated Sequencer Traces using Phred. II. Error Probabilities. *Genome Research*, (8), 175–185.

Faux, P., & Druet, T. (2017). A strategy to improve phasing of whole-genome sequenced individuals through integration of familial information from dense genotype panels. *Genetics Selection Evolution*, *49*(1), 1–13. https://doi.org/10.1186/s12711-017-0321-6

Ferry, J. A. (2006). Burkitt's Lymphoma : Clinicopathologic Features and Differential Diagnosis. *The Oncologist*, *11*, 375–383.

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., … Giuffra, E. (n.d.). Multi-species annotation of transcriptome and chromatin structure in domesticated animals, 1–25.

Fragoso, C., Heffelfinger, C., Zhao, H., & Dellaporta, S. (2016). Imputing Genotypes in Biallelic Populations from Low-Coverage Sequence Data. *Genetics*, *202*(February), 487–495. https://doi.org/10.1534/genetics.115.182071

Frischknecht, M., Pausch, H., Bapst, B., Signer-Hasler, H., Flury, C., Garrick, D., … Gredler-Grandl, B. (2017). Highly accurate sequence imputation enables precise QTL mapping in Brown Swiss cattle. *BMC Genomics*, *18*(1), 1–10. https://doi.org/10.1186/s12864-017-4390-2

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). Minimac2: Faster genotype imputation. *Bioinformatics*, *31*(5), 782–784.

https://doi.org/10.1093/bioinformatics/btu704

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.

Gengler, N., Mayeres, P., & Szydlowski, M. (2007). animal A simple method to approximate gene content in large pedigree populations : application to the myostatin gene in dual-purpose Belgian Blue cattle, 21–28. https://doi.org/10.1017/S1751731107392628

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, *51*(4), 1440. https://doi.org/10.2307/2533274

Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. *Cell*, *128*(4), 635–638. https://doi.org/10.1016/j.cell.2007.02.006

Grosclaude, F., Mahé, M.-F., Brignon, G., Di Stasio, L., & Jeunet, R. (1987). A Mendelian polymorphism underlying quantitative variations of goat αs1-casein. *Genetics Selection Evolution*, *19*(4), 399. https://doi.org/10.1186/1297-9686-19-4-399

Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W., & Van Der Werf, J. H. J. (2012). Accuracy of genotype imputation in sheep breeds. *Animal Genetics*, *43*(1), 72–80. https://doi.org/10.1111/j.1365-2052.2011.02208.x

Hayes, B. J., & Daetwyler, H. D. (2019). 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle : Applications and Outcomes.

Hayes, B. J., Macleod, I. M., Daetwyler, H. D., Bowman, P. J., Chamberlian, A. J., Vander Jagt, C. J., … Goddard, M. E. (2014). Genomic Prediction from Whole Genome Sequence in Livestock: the 1000 Bull Genomes Project. In *Proceedings, 10th World Congress of Genetics Applied to Livestock Production*.

Heidaritabar, M., Calus, M. P. L., Vereijken, A., Groenen, M. A. M., & Bastiaansen, J. W. M. (2015). Accuracy of imputation using the most common sires as reference population in layer chickens. *BMC Genetics*, *16*(1), 1–14. https://doi.org/10.1186/s12863-015-0253-5

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, *31*(2), 423–447.

Henderson, C. R., Kempthorne, O., Searle, S. R., & Krosigk, C. M. (1959). The Estimation of Environmental and Genetic Trends from Records Subject to Culling Author (s): C . R . Henderson , Oscar Kempthorne , S . R . Searle , C . M . von Krosigk Published by : International Biometric Society Stable URL : http://www.jstor.org/s. *Biometrics*, *15*(2), 192–218. https://doi.org/10.2307/2527669

Hickey, J. M., Kinghorn, B. P., Tier, B., Van Der Werf, J. H., & Cleveland, M. A. (2012). A phasing and imputation method for pedigreed populations that results in a single-stage

genomic evaluation. *Genetics Selection Evolution*, *44*(1), 1–11. https://doi.org/10.1186/1297-9686-44-9

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367. https://doi.org/10.1073/pnas.0903103106

Hozé, C., Fouilloux, M., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., … Croiseau, P. (2013). High-density marker imputation accuracy in sixteen French cattle breeds. *Genetics Selection Evolution*, 1–11.

Johnston, J., Kistemaker, G., & Sullivan, P. G. (2011). Comparison of different imputation methods. *Interbull Bull*, *44*(44).

Kanetsky, P. A., Swoyer, J., Panossian, S., Holmes, R., Guerry, D., & Rebbeck, T. R. (2002). A Polymorphism in the Agouti Signaling Protein Gene Is Associated with Human Pigmentation, 770–775.

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J. A. C., Baurain, D., … Coppieters, W. (2011). Articles Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature, *43*(5). https://doi.org/10.1038/ng.814

Koning, A. P. J. De, Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome, *7*(12). https://doi.org/10.1371/journal.pgen.1002384

Legarra, A. (2014). Bases for Genomic Prediction, 87. https://doi.org/10.1016/j.joca.2004.04.008

Legarra, A., & Vitezica, Z. G. (2015). Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using gene content multiple-trait BLUP. *Genetics Selection Evolution*, *47*(1), 89. https://doi.org/10.1186/s12711-015-0165-x

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Sargolzaei, M., & Schenkel, F. (2014). Accuracy of Whole-genome Sequence Genotype Imputation in Cattle Breeds. *World Congress on Genetics Applied to Livestock Production*. Retrieved from https://www.asas.org/docs/default-source/wcgalp-

posters/667_paper_9613_manuscript_1052_0.pdf?sfvrsn=2

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., … Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of-target mRNAs. *Nature*, *433*(7027), 769–773. https://doi.org/10.1038/nature03315

Liu, D., Shimonov, J., Primanneni, S., Lai, Y., Ahmed, T., & Seiter, K. (2007). t(8;14;18): A 3-way chromosome translocation in two patients with Burkitt's lymphoma/leukemia. *Molecular Cancer*, *6*, 3–7. https://doi.org/10.1186/1476-4598-6-35

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., … Law, M. (2012). Comparison of Next-Generation Sequencing Systems, *2012*. https://doi.org/10.1155/2012/251364

Lourenco, D. A. L., Misztal, I., Tsuruta, S., Aguilar, I., Ezra, E., Ron, M., … Weller, J. I. (2014). Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. *Journal of Dairy Science*, *97*(3), 1742–1752. https://doi.org/10.3168/jds.2013-6916

Manfredi, E., & Ådnøy, T. (2012). Génétique des caprins laitiers. *Productions Animales*, *25*(3), 233–244.

Manfredi, E., Piacere, A., Lahaye, P., & Ducrocq, V. (2001). Genetic parameters of type appraisal in Saanen and Alpine goats, *70*, 183–189.

Martin, P. (2016). *Identification et caractérisation fonctionnelle de régions associées à des caractères d'intérêt pour la filière caprine*.

Martin, P., & Leroux, C. (2000). Caprine gene specifying alpha(s1)-casein: A highly suspicious factor with both multiple and unexpected effects. *PRODUCTIONS ANIMALES*, (SI), 125–132.

Martin, P., & Leroux, C. (2000). Le gène caprin spécifiant la caséine αs1: Un suspect tout désigné aux effets aussi multiples qu'inattendus. *Productions Animales*, *October*(SPEC.ISS.), 125–132.

Martin, P. M., Palhière, I., Ricard, A., Tosser-Klopp, G., & Rupp, R. (2016). Genome wide association study identifies new loci associated with undesired coat color phenotypes in Saanen goats. *PLoS ONE*, *11*(3), 1–15. https://doi.org/10.1371/journal.pone.0152426

Martin, P., Palhière, I., Maroteau, C., Bardou, P., Canale-, K., Sarry, J., … Tosser-klopp, G. (2017). A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content, (September 2016), 1–13. https://doi.org/10.1038/s41598-017-02052-0

Martin, P., Palhière, I., Maroteau, C., Clément, V., David, I., Tosser-Klopp, G., & Rupp, R. (2018). Genome-wide association mapping for type and mammary health traits in French

dairy goats identifies a pleiotropic region on chromosome 19 in the Saanen breed. *Journal of Dairy Science*, *0*(0), 5214–5226. https://doi.org/10.3168/jds.2017-13625

Martin, P., Palhière, I., Tosser-Klopp, G., & Rupp, R. (2017). Corrigendum to "Heritability and genome-wide association mapping for supernumerary teats in French Alpine and Saanen dairy goats" (J. Dairy Sci. 99:8891–8900). *Journal of Dairy Science*, *100*(9), 7750. https://doi.org/10.3168/jds.2017-100-9-7750

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mdladla, K., Dzomba, E. F., Huson, H. J., & Muchadeyi, F. C. (2016). Population genomic structure and linkage disequilibrium analysis of South African goat breeds using genome-wide SNP data, (Campbell 2003), 471–482. https://doi.org/10.1111/age.12442

Mesbah-Uddin, M., Guldbrandtsen, B., Lund, M. S., Boichard, D., & Sahana, G. (2019). Joint imputation of whole-genome sequence variants and large chromosomal deletions in cattle. *Journal of Dairy Science*, *102*(12), 11193–11206. https://doi.org/10.3168/jds.2019-16946

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers, 240–248. https://doi.org/10.1101/gr.5681207.high-throughput

Moghaddar, N., Macleod, I. M., Duijvesteijn, N., Bolormaa, S., Khansefid, M., Swan, A. A., … van der Werf, J. H. J. (2018). Genomic evaluation based on selected variants from imputed whole-genome sequence data in Australian sheep populations. In *Proceedings of the World Congress on Genetics Applied to Livestock Production*.

Monget, P., & Reiner, A. V. (2014). *Introduction à la génétique moderne*. Les Editions de l'Ecole Polytechnique.

Mucha, S., Mrode, R., Coffey, M., Kizilaslan, M., Desire, S., & Conington, J. (2018). Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. *Journal of Dairy Science*, *101*(3), 2213–2225. https://doi.org/10.3168/jds.2017-12919

Nicoloso, L., Bomba, L., Colli, L., Negrini, R., Milanesi, M., Mazza, R., … Consortium, G. (2015). Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip. *Genetics Selection Evolution*, 1–10. https://doi.org/10.1186/s12711-015-0140-6

Oget, C., Clément, V., Palhière, I., Tosser-Klopp, G., Fabre, S., & Rupp, R. (2018). Genome-wide study finds a QTL with pleiotropic effect on semen and production traits in Saanen goats. In *Proceedings of the 69th annual meeting of the Auropean Federation of Animal*

Science.

Oget, C., Servin, B., & Palhière, I. (2019). Genetic diversity analysis of French goat populations reveals selective sweeps involved in their differentiation. *Animal Genetics*, *50*(1), 54–63. https://doi.org/10.1111/age.12752

Palhiere, I., Clement, V., Martin, P., Colleau, J. J., Palhiere, I., Clement, V., … Colleau, J. J. (2014). Bilan de la méthode de Sélection à Parenté Minimum après 6 ans d ' application dans le schéma de sélection caprin 6-year implementation of an optimization method in the breeding scheme of dairy goats, (1), 2005–2008.

Palhière, I., OGET, C., & Rupp, R. (2018). Functional longevity is heritable and controlled by a major gene in French dairy goats. *Proceedings of the World Congress on Genetics Applied to Livestock Production*, *Species-Caprine*, 165.

Pausch, H., MacLeod, I. M., Fries, R., Emmerling, R., Bowman, P. J., Daetwyler, H. D., & Goddard, M. E. (2017). Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genetics Selection Evolution*, *49*(1), 1–30. https://doi.org/10.1186/s12711-017-0301-x

Purcell, D. F., & Martin, M. A. (1993). Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *Journal of Virology*, *67*(11), 6365–6378. Retrieved from https://jvi.asm.org/content/67/11/6365

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., … Gu, Y. (2012). A tale of three next generation sequencing platforms : comparison of Ion Torrent , Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 1. https://doi.org/10.1186/1471-2164-13-341

Reber, I., Keller, I., Becker, D., Flury, C., & Welle, M. (2015). Wattles in goats are associated with the FMN1 / GREM1 region on chromosome 10, 316–320. https://doi.org/10.1111/age.12279

Rieder, S., Taourit, S., Mariat, D., & Langlois, B. (2001). Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus), *455*, 450–455. https://doi.org/10.1007/s003350020017

Robert-Granié, C., Legarra, A., & Ducrocq, V. (2011). Principes de base de la sélection génomique. *Productions Animales*, *24*(4), 331–340.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Rupp, R., Clément, V., Piacere, A., & Manfredi, E. (2011). Genetic parameters for milk

somatic cell score and relationship with production and udder type traits in dairy Alpine and Saanen primiparous goats. *Journal of Dairy Science*, *94*(7), 3629–3634. https://doi.org/10.3168/jds.2010-3694

Rupp, R., Senin, P., Sarry, J., Allain, C., Tasca, C., Ligat, L., … Tosser-Klopp, G. (2015). A Point Mutation in Suppressor of Cytokine Signalling 2 (Socs2) Increases the Susceptibility to Inflammation of the Mammary Gland while Associated with Higher Body Weight and Size and Higher Milk Production in a Sheep Model. *PLoS Genetics*, *11*(12), 1–19. https://doi.org/10.1371/journal.pgen.1005629

Sahana, G., Guldbrandtsen, B., Lund, M. S., & Vilkki, J. (2016). Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. *BMC Genetics*, 1–12. https://doi.org/10.1186/s12863-016-0363-8

Sanchez, M. P., Ferrand, M., Gelé, M., Pourchet, D., Amigues, Y., & Fritz, S. (2016). Whole-genome scan to detect quantitative trait loci associated with milk protein composition in 3 French dairy cattle breeds. *Journal of Dairy Science*, *99*(10), 8203–8215. https://doi.org/10.3168/jds.2016-11437

Sanchez, M. P., Ramayo-Caldas, Y., Wolf, V., Laithier, C., El Jabri, M., Michenet, A., … Boichard, D. (2019). Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. *Genetics Selection Evolution*, *51*(1), 1–19. https://doi.org/10.1186/s12711-019-0473-7

Sargolzaei, M., Chesnais, J. P., & Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, *15*(1). https://doi.org/10.1186/1471-2164-15-478

Simonsen, K. L., Churchill, G. A., & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, *141*(1), 413–429.

Steiner, D. F., Cunningham, D., Spigelman, L., & Aten, B. (1967). Insulin Biosynthesis: Evidence for a Precursor. *Science*, *157*(3789), 697–700. https://doi.org/10.1126/science.157.3789.697

Swarts, K., Li, H., Navarro, J. A. R., An, D., Romay, M. C., Hearne, S., … Bradbury, P. J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage , Next-Generation Sequence Data in Crop Plants, (November), 1–12. https://doi.org/10.3835/plantgenome2014.05.0023

Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays*, *29*(3), 288–299. https://doi.org/10.1002/bies.20544

Talouarn, E., Bardou, P., Palhière, I., Oget, C., Clément, V., Vargoats, T., … Robert-granié, C. (2020). Genome wide association analysis on semen volume and milk yield using

different strategies of imputation to whole genome sequence in French dairy goats, 1–13.

Teissier, M. (2019). *Intégration de données génomiques (mutations, gènes majeurs, marqueurs SNP, haplotypes) dans les modèles d'évaluations génétiques des chèvres laitières pour améliorer l'efficacité de la sélection*. Institut National Polytechnique de Toulouse.

Teissier, M., Larroque, H., & Granié, C. R. (2018). Weighted single - step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats : a quantitative trait influenced by a major gene. *Genetics Selection Evolution*, 1–12. https://doi.org/10.1186/s12711-018-0400-3

Teissier, M., Larroque, H., & Robert-Granie, C. (2019). Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats. *Journal of Dairy Science*, *102*(4), 3142–3154. https://doi.org/10.3168/jds.2018-15650

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., Dong, Y., … Zhao, S. (2014). Design and characterization of a 52K SNP chip for goats. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0086227

Tosser-Klopp, G., Bardou, P., Cabau, C., Eggen, a, Faraut, T., Heuven, H., … Zhang, W. (2012). Goat genome assembly, Availability of an international 50K SNP chip and RH panel: An update of the International Goat Genome Consortium projects. *Plant and Animal Genome Conference*, 1–14.

Vallejo, R. L., Leeds, T. D., Gao, G., Parsons, J. E., Martin, K. E., Evenhuis, J. P., … Palti, Y. (2017). Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree - based model in rainbow trout aquaculture. *Genetics Selection Evolution*, 1–13. https://doi.org/10.1186/s12711-017-0293-6

Van Binsbergen, R., Bink, M. C. A. M., Calus, M. P. L., Van Eeuwijk, F. A., Hayes, B. J., Hulsegge, I., & Veerkamp, R. F. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution*, *46*(1), 1–13. https://doi.org/10.1186/1297-9686-46-41

van Son, M., Enger, E. G., Grove, H., Ros-Freixedes, R., Kent, M. P., Lien, S., & Grindflek, E. (2017). Genome-wide association study confirm major QTL for backfat fatty acid composition on SSC14 in Duroc pigs. *BMC GENOMICS*, *18*. https://doi.org/10.1186/s12864-017-3752-0

Vanraden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., … Doak, G. A. (2013). Genomic imputation and evaluation using high-density Holstein

genotypes. *Journal of Dairy Science*, *96*(1), 668–678. https://doi.org/10.3168/jds.2012-5702

Vanraden, P. M., & Sullivan, P. G. (2010). International genomic evaluation methods for dairy cattle, 1–9.

Vanraden, P. M., Tooker, M. E., Connell, J. R. O., Cole, J. B., & Bickhart, D. M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genetics Selection Evolution*, 1–12. https://doi.org/10.1186/s12711-017-0307-4

Ventura, R. V., Lu, D., Schenkel, F. S., Wang, Z., Li, C., & Miller, S. P. (2014). Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. *Journal of Animal Science*, *92*(4), 1433–1444. https://doi.org/10.2527/jas.2013-6638

Ventura, R. V., Miller, S. P., Dodds, K. G., Auvray, B., Lee, M., Bixley, M., … McEwan, J. C. (2016). Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. *Genetics Selection Evolution*, *48*(1), 1–20. https://doi.org/10.1186/s12711-016-0244-7

Visser, C., Lashmar, S. F., Marle-köster, E. Van, & Poli, M. A. (2016). Genetic Diversity and Population Structure in South African , French and Argentinian Angora Goats from Genome-Wide SNP Data, 1–15. https://doi.org/10.5061/dryad.p1b90.

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Fernando, R. L., Vitezica, Z., … Muir, W. M. (2014). Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Frontiers in Genetics*, *5*(MAY), 1–10. https://doi.org/10.3389/fgene.2014.00134

Wang, W., Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., … Wang, J. (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nature Biotechnology*, *31*(2), 135–141. https://doi.org/10.1038/nbt.2478

Williams, E. J. (1959). The Comparison of Regression Variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 396–399. https://doi.org/10.1111/j.2517-6161.1959.tb00346.x

Wolc, a., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., … Dekkers, J. C. M. (2011). Comparison of the accuracy of genotype imputation using different methods. *7 Th European Symposium on Poultry Genetics*, *2*(October), 76. Retrieved from http://www.roslin.ed.ac.uk/7espg/assets/7espg-edited-proceedings.pdf

Ye, S., Yuan, X., Lin, X., Gao, N., Luo, Y., Chen, Z., … Zhang, Z. (2018). Imputation from SNP chip to sequence: A case study in a Chinese indigenous chicken population. *Journal of Animal Science and Biotechnology*, *9*(1), 1–12. https://doi.org/10.1186/s40104-018-0241-5

Zhang, C., Kemp, R. A., Stothard, P., Wang, Z., Boddicker, N., Krivushin, K., … Plastow, G. (2018). Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K , 650K and whole - genome sequence variants. *Genetics Selection Evolution*, 1–13. https://doi.org/10.1186/s12711-018-0387-9

Zhang, X., Lourenco, D., Aguilar, I., & Legarra, A. (2016). Weighting Strategies for Single-Step Genomic BLUP : An Iterative Approach for Accurate Calculation of GEBV and GWAS, *7*(August), 1–14. https://doi.org/10.3389/fgene.2016.00151

Zheng, C., Boer, M. P., & Eeuwijk, F. A. Van. (2018). Accurate Genotype Imputation in Multiparental Populations from Low-Coverage Sequence, *210*(September), 71–82.

Résumé grand public

La démocratisation des données de séquençage tout génome pour les animaux de rente ouvre de nouvelles perspectives pour la sélection. Le projet VarGoats a pour but de mettre à disposition d'un consortium international, un jeu de données de plus de 1000 séquences pour l'espèce *Capra hircus*. Ces premières données dans l'espèce caprine nous ont permis d'étudier les modalités de vérification de la qualité des séquences et d'exploiter cette information dans les programmes d'amélioration génétique des caprins laitiers français. Dans cette thèse, nous avons montré que les informations de séquence permettent de mieux comprendre l'architecture génétique de nombreux caractères d'intérêt pour la filière, d'améliorer la précision des évaluations génomiques et ainsi d'optimiser la sélection en races Alpine et Saanen françaises. La thèse a été soutenue financièrement par la région Occitanie et le département Génétique Animale INRAE.

General public abstract

The recent availability of whole-sequence data for farm animals opens up new perspectives regarding selection. The VarGoats project is an international 1,000 genomes resequencing program designed to provide sequence information of the *Capra hircus* species. The first sequenced individuals of the species allowed us to implement a solid quality check and take advantage of sequence data in the breeding programs of French dairy goats. In this thesis applied to French dairy goats, we showed that sequence information provides better insight on the genetic architecture of traits of interest for the breeders and significantly increases the precision of genomic evaluations in French Alpine and Saanen breeds. This thesis was financed by the Occitanie region and the Animal Genetics division of INRAE.

Résumé

La filière caprine française a intégré l'ère de la génomique avec le récent développement et la valorisation d'une puce à ADN dans les années 2010-2020 pour la recherche de QTL et l'évaluation génétique. La démocratisation des données de séquençage tout génome pour les animaux de rente ouvre de nouvelles perspectives. Le projet VarGoats, a pour but de mettre à disposition d'un consortium international, un jeu de données de plus de 1000 séquences pour l'espèce *Capra hircus*. L'étude de la qualité d'imputation vers la séquence dans la filière caprine est un préalable nécessaire à l'utilisation de cette dernière dans les analyses d'association pour la détection de QTL ainsi que dans les évaluations génomiques. L'objectif principal de ces travaux est d'étudier l'intégration potentielle des données de séquence dans les programmes d'amélioration génétique de la filière laitière caprine française. La mise en place d'un contrôle de la qualité des données de séquence a représenté un travail majeur dans ma thèse. Il s'est appuyé sur une recherche bibliographique ainsi que sur la comparaison des génotypes 50k disponibles avec les séquences filtrées. Finalement, sur les 97 889 899 SNP et 12 304 043 indels initiaux, nous avons retenu 23 338 436 variants dont 40 491 appartenaient au set de SNP de la puce Illumina GoatSNP50 BeadChip.

Une étude préalable de l'imputation depuis la puce 50k vers la séquence a ensuite été menée dans le but d'obtenir un nombre suffisant de séquences imputées de bonne qualité. Plusieurs méthodes d'imputation (imputation populationnelle ou familiale) et plusieurs logiciels ont été testés en utilisant les données de séquence disponibles (829 séquences des différences races caprines internationales). En intra-race, les taux de concordances génotypiques et alléliques ont été estimées à 0,74 et 0,86 en Saanen et 0,76 et 0,87 en Alpine respectivement. Les corrélations étaient alors de 0,26 et 0,24 en Alpine et Saanen respectivement.

Les séquences imputées des mâles ont permis la confirmation de QTL précédemment observés sur les génotypes 50k ainsi que la détection de nouvelles régions d'intérêt. L'exhaustivité des données de séquence représentait une opportunité sans précédent d'approfondir une région QTL du chromosome 19 en Saanen qui est associée à la fois à des caractères de production mais aussi à des caractères de morphologie et santé de la mamelle ainsi qu'à des caractères de production de semence. Cette analyse n'a pas abouti à l'identification de mutations candidates. Néanmoins, nous avons pu proposer un moyen simple d'identifier des profils génomiques et phénotypiques particuliers en race Saanen à partir d'un génotype 50k. Cette méthode pourra s'avérer utile en terme de prédiction précoce tant en France qu'à l'international.

Enfin, en réunissant l'ensemble des travaux effectués précédemment, nous avons étudié l'impact de l'intégration de données de séquence imputées sur le chromosome 19 sur la précision des évaluations en race Saanen françaises. Plusieurs modèles d'évaluations ont été mis en œuvre et comparés : single-step GBLUP (ssGBLUP), single-step GBLUP pondéré (WssGBLUP) en utilisant différents panels de variants imputés. Les meilleurs résultats ont été obtenus en utilisant un ssGBLUP incluant les génotypages 50k et les variants imputés de la région du QTL du chromosome 19 (entre 24,72 et 28,38 Mb) avec des gains de +6,2% de précision en moyenne sur les caractères évalués. La mise à jour de la puce caprine à laquelle j'ai participé représente une perspective d'amélioration de la précision des évaluations. Elle permet d'améliorer significativement la qualité des évaluations génomiques (entre 3,1 et 6,4% en fonction du scenario considéré) tout en limitant les temps de calculs liés à l'imputation notamment. Ces travaux confortent l'intérêt de l'utilisation de données de séquence dans les programmes de sélection caprins français et ouvrent la perspective de leur intégration dans la routine des évaluations.

Abstract

French dairy goats recently integrated genomics with the development of a DNA chip in the 2010s and the first QTL detections and genomic evaluations. The availability of sequence data for farm animals opens up new opportunities. The VarGoats project is an international 1,000 genomes resequencing program designed to provide sequence information of the *Capra hircus* species. The study of imputation quality to sequence level is a necessary first step before using imputed sequences in association analysis and genomic evaluations. The main objective of this work was to study the possible integration of sequence data in the French dairy goats breeding programs. The set up of a quality check represented a sizable part of this thesis. It was based on bibliographic research and the comparison between available 50k genotypes and sequence data. Out of the initial 97,889,899 SNPs and 12,304,043 indels, we eventually retained 23,338,436 variants including 40,491 SNPs of the Illumina GoatSNP50 BeadChip.

A preliminary study of imputation from 50k genotypes to sequence was then performed with the aim of getting a sufficient number of sequenced animals of good quality. Several softwares and methods were considered (family or population imputation) using the 829 sequenced animals available. Within-breed imputation led to genotype and allele concordance of 0.74 and 0.86 in Saanen and 0.76 and 0.87 in Alpine respectively. Correlations were then of 0.26 and 0.24 in Alpine and Saanen respectively.

Imputed sequence of males confirmed signals previously identified using 50k genotypes and allowed the detection of new regions of interest. The density of sequence data represented an unprecedented opportunity to deepen our understanding of QTL region of chromosome 19 in the Saanen breed. This region is associated to production, type and udder health traits as well as semen production traits. Our analysis did not point out any candidate mutation. However, we offer a simple way to identify genomic and phenotypic profiles in the Saanen breed using 50k genotypes. This method could be of use for early prediction in France but also worldwide.

Finally, using all previous results, we studied the impact of the integrating imputed sequence data of chromosome 19 on the accuracy of evaluations in French Saanen. Several evaluation models were compared : single-step GBLUP (ssGBLUP) and weighted single-step GBLUP (WssGBLUP) using different panels of imputed variants. Best results were obtained using ssGBLUP with 50k genotypes and all variants on the QTL region of chromosome 19 (between 24.72 and 28.38Mb): +6.2% accuracy on average for all evaluated traits. The 50k chip update to which I participated represents a opportunity to improve genomic evaluations. Indeed, it significantly improved accuracy of predictions (between 3.1 and 6.4% on average depending on the scenario) while limiting computation time associated to imputation. This work confirms the benefits of using sequence data in the French dairy goats breeding programs and opens up the perspective of integrating them in the routine genomic evaluations.